Identification of cancer driver genes using supervised machine learning and systems biology

Mourikis, Athanasios

Awarding institution:
King's College London

Identification of cancer driver genes using supervised machine learning and systems biology

Thanos Mourikis

A thesis submitted for the degree of Doctor of Philosophy

King's College London

January 2019

Declaration

I, Thanos Mourikis, confirm that the work presented in this thesis is my own and has not been submitted in any form for another degree or diploma at any university or other institution or tertiary education. Information that has been derived from the work of others is clearly indicated and attributed.

Abstract

During the past few years many tumour sequencing projects have been focused on the characterisation of genes harbouring somatic alterations with cancer promoting role, which have been named cancer driver genes. Since these genes have been shown to be subject to positive selection during cancer progression, it has been assumed that their mutation is observed more frequently than expected. However, the full characterisation of cancer drivers is particularly challenging in cancer types, such as oesophageal adenocarcinoma (OAC), in which the genomic landscape is highly variable and recurrent events are not frequent.

To identify rare or even patient-specific cancer driver genes, a novel algorithm, sysSVM, was developed. SysSVM is based on support vector machines, a supervised machine-learning framework, and utilises systems-level properties of human genes and sequencing data from individual tumours to predict genes that promote cancer development. Unlike other state-of-the-art algorithms for driver gene prediction, sysSVM takes into account all types of damaging alterations simultaneously (mutations, copy number alterations and structural rearrangements). After the development phase, sysSVM was applied to 261 OACs from the Oesophageal Cancer Clinical and Molecular Stratification (OCCAMS) consortium. A large number of novel cancer driver genes that, together with well-known drivers, help promote OAC was discovered. Validation of sysSVM using 107 additional OACs confirmed the robustness of the approach. Moreover, the large majority of the newly discovered cancer genes was rare or patient-specific. Despite this, it was shown that they converged towards perturbing the same cancer-related processes, including intracellular signalling and cell cycle regulation. Recurrence of process perturbation, rather

than mutations in individual genes, divided OACs into six clusters that differ in their molecular and clinical features, suggesting patient stratifications for personalised treatments. Collaboration with bench researchers to experimentally mimic or reverting alterations of the predicted cancer driver genes, validated their contribution to cancer progression in OAC.

The findings of this thesis accomplish three things. First, they describe the first attempt to develop an algorithm, which extends the discovery of somatically acquired perturbations contributing to cancer beyond those of recurrent driver genes. Second, they reveal a widespread somatic perturbation of biological processes in OAC, demonstrate OAC acquired dependencies and highlight potential therapeutic targets. Third, they provide insights into the potential use of the newly predicted cancer driver genes to stratify OACs and inform clinical practice.

Acknowledgements

I would like to thank my primary supervisor Prof. Francesca Ciccarelli, as well as the members of my thesis committee, Dr. Anita Grigoriadis, Prof. Franca Fraternali, Prof. Michael Simpson and Dr. Sophia Tsoka. Special thanks go to Dr. Christopher Yau for his guidance and support throughout my thesis. Thanks should also go to members of the Cancer Systems Biology Laboratory past and present. I am also grateful to external collaborators for their contribution to data and ideas, patients for consenting to the use of tissue samples from which genomic data have been generated, and funding bodies that have provided financial support for the work presented in this thesis. I would also like to give special thanks to my friends and family, who have supported me throughout my PhD. Finally, I would like to thank my wife, Kiki, for always helping me think clearly, for helping me find the answers to my questions, and for enduring endless hours of my coding.

Table of contents

List of Figures

List of Tables

Abbreviations

ACC	Adrenocortical carcinoma
ALL	Acute lympoblastic leukaemia
ASCAT	Allele-Specific Copy number Analysis of Tumours
BE	Barret's oesophagus
BMR	Background mutation rate
BRCA	Breast invasive carcinoma
CDD	Conserved Domains Database
CDK	Cyclin-dependent kinase
CGC	Cancer Gene Census
CHOL	Cholangiocarcinoma
CML	Chronic myeloid leukaemia
CNV	Copy number variation
COAD	Colon adenocarcinoma
DDK	Dbf4-dependent kinase
Edu	5-ethynyl-2'-deoxyuridine
ESCA	Oesophageal carcinoma
FPKM	Fragments per kilobase million
GBM	Glioblastoma
GEF	Guanine exchange factor
GERD	Gastroesophageal reflux disease
GOJ	Gastro-Oesophageal Junction
GTEx	Genotype-Tissue Expression
HAT	Histone acetyl-transferase
HNSC	Head and neck squamous cell carcinoma
ICGC	International Cancer Genome Consortium
InDels	Insertion-Deletion mutations
k-NN	k-Nearest Neighbour
KICH	Kidney Chromophobe
KIRC	Kidney Renal Clear Cell Carcinoma
KIRP	Kidney renal papillary cell carcinoma
KO	Knock-out
LAML	Acute myeloid leukaemia
LGG	Lower-grade glioma

LUSC	Lung squamous cell carcinoma
MEMo	Mutual Exclusivity Modules
MuSiC	Mutational Significance in Cancer
MutSigCV	Mutation Significance with Covariates
MutSigDB	Mutation Significance Database
MVE	Minimum Volume Ellipsoid
OAC	Oesophageal adenocarcinoma
OCCAMS	Oesophageal cancer clinical and molecular stratification
OSCC	Oesophageal squamous cell carcinoma
OV	Ovarian serous cystadenocarcinoma
PAAD	Pancreatic adenocarcinoma
PIN	Protein-protein interaction network
PRAD	Prostate adenocarcinoma
Rb	Retinoblastoma gene
READ	Rectum adenocarcinoma
RFE	Recursive Feature Elimination
RISC	RNA-induced silencing complex
SNV	Single nucleotide variant
SV	Structural variant
SVMs	Support Vector Machines
SysSVM	Systems-level Support Vector Machines
t-SNE	t-Distributed stochastic neighbour embedding
TCGA	The Cancer Genome Atlas
TGCT	Testicular germ cell tumours
THCA	Thyroid carcinoma
THYM	Thymoma
TLR	Toll-like receptor
UCS	Uterine carcinoma
UVM	Uveal Melanoma
WGS	Whole-genome sequencing

Chapter 1. Introduction

1.1 Cancer genome and cancer drivers

Cancer is a disease of the genome and its pathogenesis lies in the accumulation of changes in the genome of somatic cells (Stratton, Campbell, and Futreal 2009). These changes lead to widespread deregulation of cell functions, which in turn provides tumours with sufficient diversity to adapt and proliferate in the ever-changing tumour microenvironment. Depending on the cancer type, cancer genomes can be genomically stable, with very few changes present, or highly unstable, exhibiting numerous genomic changes (Jefford and Irminger-Finger 2006; Lengauer, Kinzler, and Vogelstein 1998; Vogelstein et al. 2013). These somatic alterations can affect the DNA sequence at various extents and thus, range from point mutations to large chromosomal abnormalities. Specifically, they are further divided in: (i) single nucleotide variants (SNVs), (ii) insertions/deletions of small or large DNA sequences (indels), (iii) DNA rearrangements, in which DNA segments are relocating from one genomic position to another via recombination or gene conversion, and (iv) copy number variations, which can be either increases, leading to gene amplification, or reductions, leading to removal of a DNA segment from the genome (Lawrence et al. 2013; Weir, Zhao, and Meyerson 2004; Stephens et al. 2009; Alexandrov et al. 2013). However, it should be noted that somatic changes that are critical for cancer initiation and progression are not merely genomic, rather they can be present in virtually every level of genome's organisation (Network et al. 2013; Dulak et al. 2013; Jusakul et al. 2017; Sabarinathan et al. 2017). For example, epigenetic changes, such as DNA methylation of cytosine residues, can regulate gene expression and thus,

profoundly affect cancer onset and progression (Sharma, Kelly, and Jones 2010).

The compendium of alterations in cancer genomes is not fixed, but changes over time as tumours evolve and this evolutionary nature of neoplasms has been described since 1976 (Nowell 1976). Mutational patterns in cancer are shaped by external factors, such as exposure to cigarette carcinogens and UV light, or may be introduced by the cell's own error-prone DNA repair mechanisms (Stratton 2011). Although randomly distributed mutations accumulate in normal cells even before cell transformation, they can be subsequently subject to positive selection. Therefore, at any point in time, the mutational landscape of cancer genome reflects the outcome of random events and selection of mutations that confer a fitness advantage to tumour cells (Martincorena et al. 2017, 2015; Tomasetti, Vogelstein, and Parmigiani 2012). Thus, cancer evolution has been characterised as a typical Darwinian system, in which selective pressures act on genetic alterations that trigger phenotypic changes (Greaves and Maley 2012; Cahill et al. 1999). The study of tumours from an evolutionary perspective is critical for the understanding of tumorigenesis and therapeutic resistance arising after treatment (Meads, Gatenby, and Dalton 2009; Moulder 2010). As tumour evolves, the set of somatic changes in cancer cells is continuously shaped by a combination of selective pressures and neutrality. Consequently, somatic alterations can be classified into: i) those that confer a selective advantage to the cell, called "driver" alterations, ii) those that are detrimental for cell's fitness and thus, subject to negative selection, and iii) those that are selectively neutral with limited or no contribution to tumorigenesis and therefore termed "passenger"

alterations (Pon and Marra 2015; Haber and Settleman 2007; Maley et al. 2004; Simpson 2009).

More recently, it was suggested that in addition to these three types of alterations there is another intermediate form that lies between drivers and passengers. These alterations have been termed “mini drivers” and they exemplify an alleviated form of major drivers (Li and Thirumalai 2016; Castro-Giner, Ratcliffe, and Tomlinson 2015). According to the mini driver hypothesis, in tumours lacking a major driver, oncogenic transformation may occur through a combined action of multiple mini-driver genes, each with a weak individual effect. Their potential role in tumorigenesis and disease progression will be discussed in more detail at the end of this chapter.

Understanding the role that genetic modifications play in cancer and identifying driver mutations and the cancer genes that they alter have been the main focus of cancer research. The first findings came from observational studies in the beginning of the 20th century, which highlighted that abnormalities in the hereditary material of the cells was the cause for the cancerous cell phenotype (von Hansemann 1890; Boveri 1914). Shortly after the discovery of DNA, the development of cytogenetic methods led to association of a translocation between chromosomes 9 and 22, which is known as “Philadelphia” translocation, to chronic myeloid leukaemia (Rowley 1973). Apart from chromosomal changes, many viral agents were also linked to various cancer types. Prominent examples are Rous sarcoma virus (Rous 1910), Epstein-Barr virus, which has been linked to non-Hodgkin Lymphoma (Teras et al. 2015), and long-standing infections of hepatitis B or C viruses, which have been shown to lead to liver cancer (Koike et al. 2008). These discoveries advanced our understanding of tumorigenesis at a cellular level, but a detailed

understanding of the mode of action of cancer-causing agents and the molecular, tumour-promoting changes that they introduce at a subcellular level was lacking.

Later on, several studies in the field of chemical carcinogenesis provided evidence that carcinogens deregulate cell proliferation by forming covalent bonds with cellular macromolecules, but their exact target was not identified (Miller and Miller 1947; Cook, Hewett, and Hieger 1933). The discovery of DNA as the genetic material and the description of its structure (Watson and Crick 1953) made apparent that DNA is the macromolecule that is targeted by carcinogens. Therefore, understanding the effects of individual DNA mutations was the key in dissecting basic cancer mechanisms (Loeb and Harris 2008). Consequently, an increasing interest in characterising such mutations focused on the isolation of specific DNA segments responsible for tumorigenesis. These efforts eventually led to the seminal discovery of the first cancer-causing sequence change, the substitution of glycine to valine in codon 12 of the *HRAS* gene (Reddy et al. 1982; Stratton, Campbell, and Futreal 2009).

Most driver alterations are associated with genes (also known as driver genes) that in normal cells are involved in pathways related to cellular proliferation, apoptosis and cell signalling. Based on their function, these genes can be broadly divided into two major classes – tumour suppressors and oncogenes (Futreal et al. 2004). Tumour suppressors act to repress the potential of uncontrolled proliferation of cells and their driver role in cancer is the result of inactivating mutations, whereas oncogenes result from the activating mutations of proto-oncogenes. The first tumour suppressor gene was discovered following the description of the “two-hit” hypothesis in 1971 by Alfred Knudson (Knudson Jr. 1971). He was the first to suggest that the development

of retinoblastoma, a rare childhood eye tumour, requires two damaging mutations, and as further studies confirmed, both of these mutations must occur in the two functional copies of the retinoblastoma gene (*Rb*) (Sparkes et al. 1980, 1983; Horowitz et al. 1990).

Subsequent discoveries highlighted the role of another tumour suppressor gene, *TP53*, which was shown to be the key regulator of cell proliferation (Lane and Crawford 1979). *TP53* is the most frequently mutated gene in human cancers and its role in the control of cell cycle (Levine 1997) renders it a primary target of inactivating, usually missense, mutations in cancer. Furthermore, inheritance of *TP53* mutations predisposes to multiple cancers, such as breast carcinomas, sarcomas and brain tumours (Olivier, Hollstein, and Hainaut 2010). However, *TP53* may also harbour gain-of-function mutations with oncogenic potential that make it a potential therapeutic target (Soussi and Wiman 2015). The discovery of the first tumour suppressor genes was followed by the identification of several oncogenes that were initially described in tumour-inducing retroviruses (Vogt 2012). The *src* gene in Rous sarcoma virus was the first oncogene to be characterised, due to its ability to oncogenically transform normal cells by altering cell morphology, adhesion, motility, survival, and proliferation (Martin 1970; Hunter and Sefton 1980). Many human oncogenes, such as *MYC* and *RAS*, followed the discovery of *src* and they have been subsequently recognised as critical driving forces in many types of cancer (Dang 2012; Little et al. 2011; Misale et al. 2012).

1.2 Computational algorithms to identify cancer driver genes

The discovery of driver genes was accelerated by the completion of the sequencing of the human genome (International Human Genome Sequencing Consortium 2004) and the subsequent advent of next generation sequencing technologies (Meyerson, Gabriel, and Getz 2010). With the base-level resolution of human genome available, many ambitious cancer sequencing projects were initiated, yielding sequencing data from thousands of cancer genomes (International Cancer Genome Consortium et al. 2010). Ultimately, owing to the advances in sequencing technologies and the generation of sequencing data, cancer genomics research inevitably became tightly connected to computer science and mathematics. The development of new algorithms aiming to interrogate sequencing data for somatic and germline alterations enabled a number of studies to unravel the complexity of cancer genomes in detail (Tian, Basu, and Capriotti 2015; Chin et al. 2011).

Algorithms to identify driver genes amongst all mutated genes in a cohort of cancer samples can be conceptually divided into three groups: (i) those that predict the effect of the acquired mutations on the encoded protein, (ii) those that measure the background mutation rate of the cancer genome and identify genes deviating from it and (iii) those that utilize systems-biology data (mostly protein-protein interaction networks) to assess mutation significance (Ding et al. 2014).

The first group of methods relies on the hypothesis that drivers alter the function of the normal protein either by disrupting it (loss-of-function), or by enhancing it (gain-of-function). Such methods utilise genome annotation databases, for instance Ensembl or that of University of California Santa Cruz, and predict the effect of mutations based on:

- the functional impact on the encoded protein: SIFT (Kumar, Henikoff, and Ng 2009), PolyPhen-2 (Adzhubei et al. 2010), MutationTaster (Schwarz et al. 2010), MutationAssessor (Reva, Antipin, and Sander 2011),
- sequence conservation across the tree of life: PhyloP (Pollard et al. 2010), GERP++ RS (Davydov et al. 2010), SiPhy (Garber et al. 2009), or
- mutation clustering: OncodriveClust (Tamborero, Gonzalez-Perez, and Lopez-Bigas 2013).

To extend these methods, it has been hypothesised that genes exhibiting a bias towards accumulation of mutations with high functional impact (FM bias) may be subject to positive selection. Thus, an estimation of this bias can be used to detect candidate driver genes (Gonzalez-Perez and Lopez-Bigas 2012). Most importantly, the estimation of FM bias was one of the first efforts to overcome the limitations of recurrence-based methods (see below) and predict candidate driver genes in individual samples. This was feasible due to the premise that genes mutated in only a handful of samples may show high FM bias if the mutations they harbour are all highly deleterious. Bias estimation has not been assessed only by examining the functional impact of mutations, but also by estimating clustering of mutations around certain protein residues. The “20/20” rule first published by Vogelstein and colleagues (Vogelstein et al. 2013), and later refined by the addition of a machine-learning-based algorithm (Tokheim et al. 2016), suggested that a gene should be classified as tumour suppressor if at least 20% of its mutations are inactivating. Conversely, genes with at least 20% missense mutations around a protein residue can be classified as oncogenes. In contrast to the evaluation of FM bias, the 20/20 rule cannot predict driver mutations in genes altered in very few samples, due to limitations in statistical power.

The second, and most widely used, group of methods employs an estimation of the expected number of mutations across different regions of the cancer genome, termed background mutations rate (BMR). To identify driver genes, these methods apply a statistical model to discover genes that carry more mutations that one would expect by chance, after correction for multiple testing (Getz et al. 2007; Sjoblom et al. 2006; Lawrence et al. 2013). These recurrence-based methods are based on the hypothesis that important mutations for the development of cancer recur across samples. Therefore, the prevalence of mutation of a gene is a sign of functional selection. Recurrence of mutations is, however, influenced by many factors, such as the gene length, BMR and gene replication time, with many recent approaches trying to account for these factors as well. For example, Mutational Significance in Cancer (MuSiC) implements a multidimensional statistical evaluation to identify significantly mutated genes across a given cohort of samples (Dees et al. 2012). Moreover, recently developed methods started taking into account the non-uniform mutation rate across the genome. For instance, the latest version of Mutation Significance with Covariates (MutSigCV) calculates a gene-specific background mutation rate by incorporating expression levels and gene replication time (Lawrence et al., 2013). Despite their widespread usage, recurrence-based methods still have limitations; they fail to detect rare driver genes, such as those whose mutation rate is below the background mutation rate (Lawrence et al. 2014).

Finally, the third group of methods builds on the idea that mutations target driver genes that are associated with a relatively small number of regulatory and signalling networks. These methods employ a rigorous statistical framework to identify *de* novo mutated subnetworks within the human protein-protein

interaction network (PIN) (Vandin, Upfal, and Raphael 2011). HotNet and its successor HotNet2 (Leiserson et al. 2015) are the most representative examples of such algorithms and both use a heat diffusion model to identify significantly mutated subnetworks. Specifically, mutated genes are assigned an initial heat value, based on their mutation frequency in the cohort of interest. This heat value successively fuses to their interactors through the edges of the network, thus yielding significantly “hot” subnetworks. Heat diffusion models can potentially identify genes mutated in low frequencies based on their proximity to recurrently mutated genes. However, they fail to coherently explore parts of the network containing only infrequent mutated genes as the initial heat in these regions will be low. Another network-based approach is the Mutual Exclusivity Modules (MEMo) (Ciriello et al. 2012), which is based on mutual exclusivity of mutations in given parts of the network. MEMo first identifies highly connected proteins and then tests whether sub-networks that include such proteins show mutually exclusive mutations.

Apart from rigorous prediction algorithms, many studies during the last decade sought to investigate the drivers in several cancer types by simply ranking mutated genes in cancer samples by frequency of mutations, assuming that recurrence of mutations is indicative of significance. In fact, recurrence-based driver discovery was the most commonly used method for driver identification, accounting for 44.7% of the total 188 mutational screenings that were reported as part of the Network of Cancer Genes database (An et al. 2016) (Figure 1.1A). However, the number of recurrently mutated genes in cancer cohorts is far lower than the total number of mutated genes. As a result, the mutational landscape in cancer is comprised of a few mountains, *i.e.* genes mutated in high frequencies (such as known driver genes), and numerous hills,

i.e. genes mutated in low frequencies, as depicted in Figure 1.1B. Methods that rely on recurrence (or a refined version of it) of mutations are usually incapable of identifying hills in cancer cohorts. As a consequence, identification of cancer genes that are mutated at frequencies ≥ 20% is nearing saturation, while the discovery rate of genes that are mutated at lower frequencies is in steep increase (Lawrence et al. 2014) due to the development of novel methods (Figure 1.1C). To date, more than 3% of human genes have been associated with cancer with approximately 90% of them being somatically mutated and 20% of them being altered in germline (Simon A. Forbes et al. 2017). In addition to those known driver genes, approximately another 1000 genes, usually referred to as candidate driver genes, have been implicated in multiple cancer types with their experimental validation pending further research (An et al. 2016).

The distribution of driver genes, both known and candidate, is not uniform, leaving certain cancer types with few drivers and others with hundreds. This impacts on the number of samples whose mutations can be associated with drivers. When cancer-type-specific cancer genes (collective term for known and candidate driver genes) are considered, approximately 30% of cancer samples across all cancer types remain without any driver genes altered (Figure 1.1D). To remedy this, careful curation of known and candidate driver coupled with transfer of knowledge of known cancer genes across cancer types has been very recently employed in an effort to refine drivers in a sample-specific manner (Sabarinathan et al. 2017). However, even with careful manual annotation of cancer genes in individual samples, the number of samples without driver genes can be as high as 50% in certain cancer types with the pan-cancer percentage of samples with zero drivers being around 10% (Figure 1.1E).

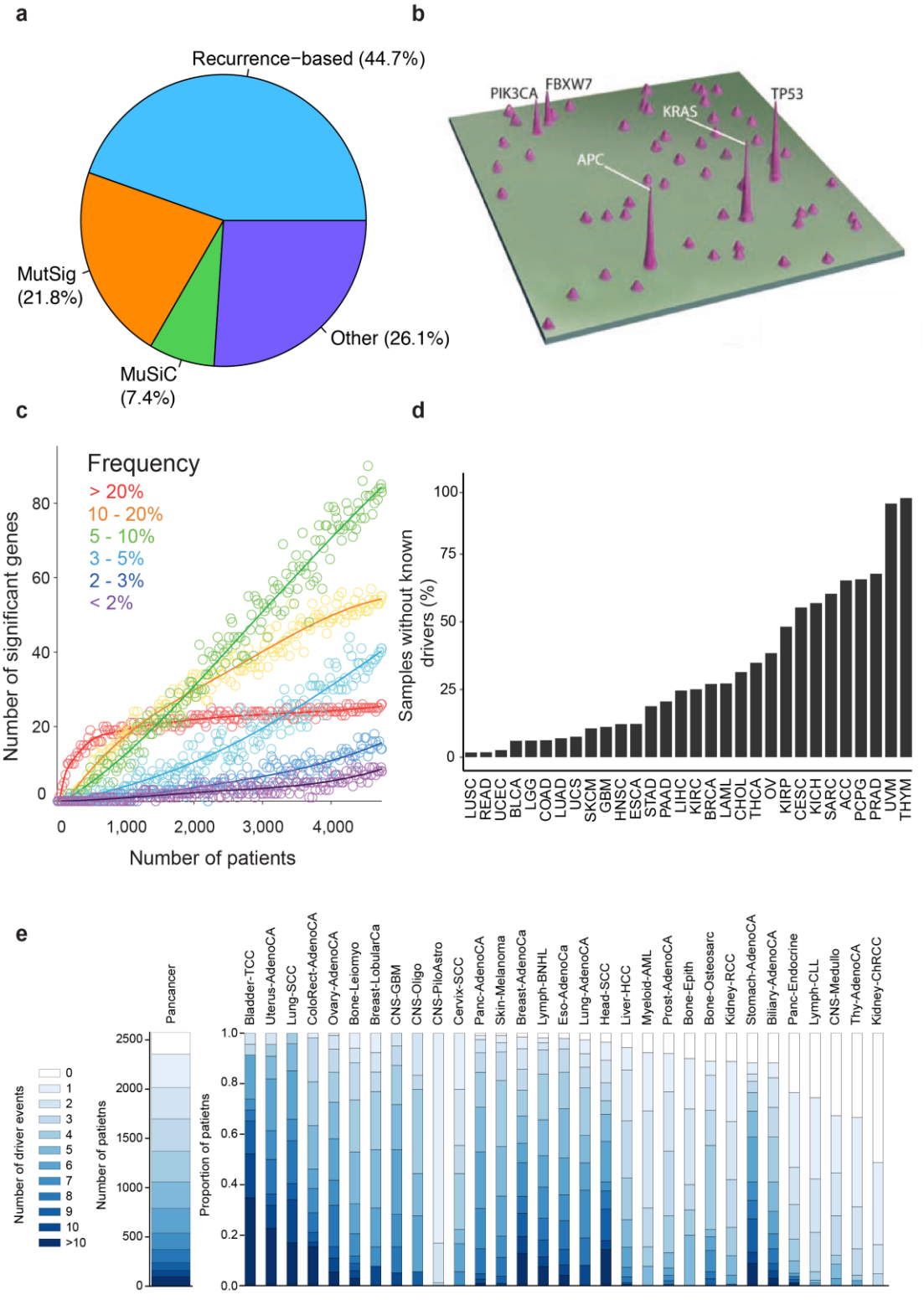


Figure 1.1. Recurrence-based driver gene identification. **(a)** Overview of the usage of methods to identify driver genes from the Network of Cancer Genes database (An et al. 2016). Pie chart was generated by the manual curation of 188 mutational screening. Figure adapted from An et al. (An et al. 2016). **(b)** Two-dimensional map of genes mutated in colorectal cancers showing a few

gene “mountains” present in a large proportion of tumours and numerous gene “hills” that are mutated infrequently. Figure taken from Wood et al. (Wood et al. 2007). **(c)** Down-sampling analysis of significant genes as a function of cohort size using MutSig. Genes were stratified by their mutation frequency in 4,742 cancer patients. Each point is a random subset of the 4,742 patients. Figure taken from Lawrence et al. (Lawrence et al. 2014). **(d)** Percentage of samples without known driver genes. Starting from 7,828 cancer samples from TCGA, all known and candidate driver genes with putative driver mutations (i.e. damaging, truncating and gain-of-function mutations, homozygous deletions, heterozygous deletions followed by damaging/truncating mutations in the second allele and amplifications) were extracted. Known and candidate driver genes were retrieved from the Network of Cancer Genes database (An et al. 2016) and associated with a specific cancer type. Full names of cancer types are reported in Table 1.1. **(e)** Stacked bar plot of tumours for 31 cancer types showing driver genes per sample. The heatmap represents the number of drivers. In contrast to figure 1d where cancer type-specific driver genes were used, in this analysis all known driver genes were considered for all samples regardless of which cancer type they had been discovered. Figure taken from Sabarinathan et al. (Sabarinathan et al. 2017).

Table 1.1. Pan-cancer cohort of 7,828 samples from The Cancer Genome Atlas consortium. For each cancer type, the number of samples is shown.

Cancer type	Abbreviation	Samples (n)
Adrenocortical Carcinoma	ACC	72
Bladder Urothelial Carcinoma	BLCA	232
Breast Invasive Carcinoma	BRCA	954
Cervical squamous cell carcinoma and endocervical adenocarcinoma	CESC	179
Cholangiocarcinoma	CHOL	35
Colon adenocarcinoma	COAD	255
Esophageal carcinoma	ESCA	179
Glioblastoma multiforme	GBM	143
Head and Neck squamous cell carcinoma	HNSC	491
Kidney Chromophobe	KICH	65
Kidney renal clear cell carcinoma	KIRC	423
Kidney renal papillary cell carcinoma	KIRP	164
Acute Myeloid Leukemia	LAML	169
Brain Lower Grade Glioma	LGG	506
Liver hepatocellular carcinoma	LIHC	187
Lung adenocarcinoma	LUAD	487
Lung squamous cell carcinoma	LUSC	174
Ovarian serous cystadenocarcinoma	OV	341
Pancreatic adenocarcinoma	PAAD	136
Pheochromocytoma and Paraganglioma	PCPG	175

Prostate adenocarcinoma	PRAD	409
Rectum adenocarcinoma	READ	110
Sarcoma	SARC	237
Skin cutaneous melanoma	SKCM	357
Stomach adenocarcinoma	STAD	335
Testicular Germ Cell Tumors	TGCT	143
Thyroid carcinoma	THCA	393
Thymoma	THYM	116
Uterine Corpus Endometrial Carcinoma	UCEC	229
Uterine Carcinosarcoma	UCS	53
Uveal melanoma	UVM	79
Total	**-**	**7,828**

Comparison and evaluation of the above-mentioned prediction algorithms and others, carried out by Tokheim and colleagues, highlighted that different approaches capture different aspects of the cancer mutational landscape and not always predict the same set of driver genes (Tokheim et al. 2016). Not surprisingly, the number of driver genes that were predicted by the various algorithms in this study ranged from approximately 200 to 2,500. Of note, methods directed towards the discovery of rare driver genes, such as OncodriveFM, identified a higher number of driver genes, as shown in figure 1.2A. Moreover, using Cancer Gene Census as reference, known cancer driver genes were identified from the output of each algorithm and their frequency within the total of predicted genes was calculated (Figure 1.2B). Overall, there was a wide range of estimated fractions of known cancer genes across the different prediction tools; for certain methods 50% of their predictions were found to be known drivers, while for others this fraction was as low as 5%. Finally, in a similar analysis An et al. (An et al. 2016) showed high overlap of predicted driver genes with known drivers, but poor performance of the same methods towards candidate driver genes (Figure 1.2C), denoting an optimisation of most methods towards known drivers. Taken together these

observations suggest that current computational approaches aiming to identify driver genes reach a consensus gene set for highly recurrent drivers. However, they lack resolution for genes that are not frequently mutated in cancer cohorts, some of them with validated tumorigenic effect.

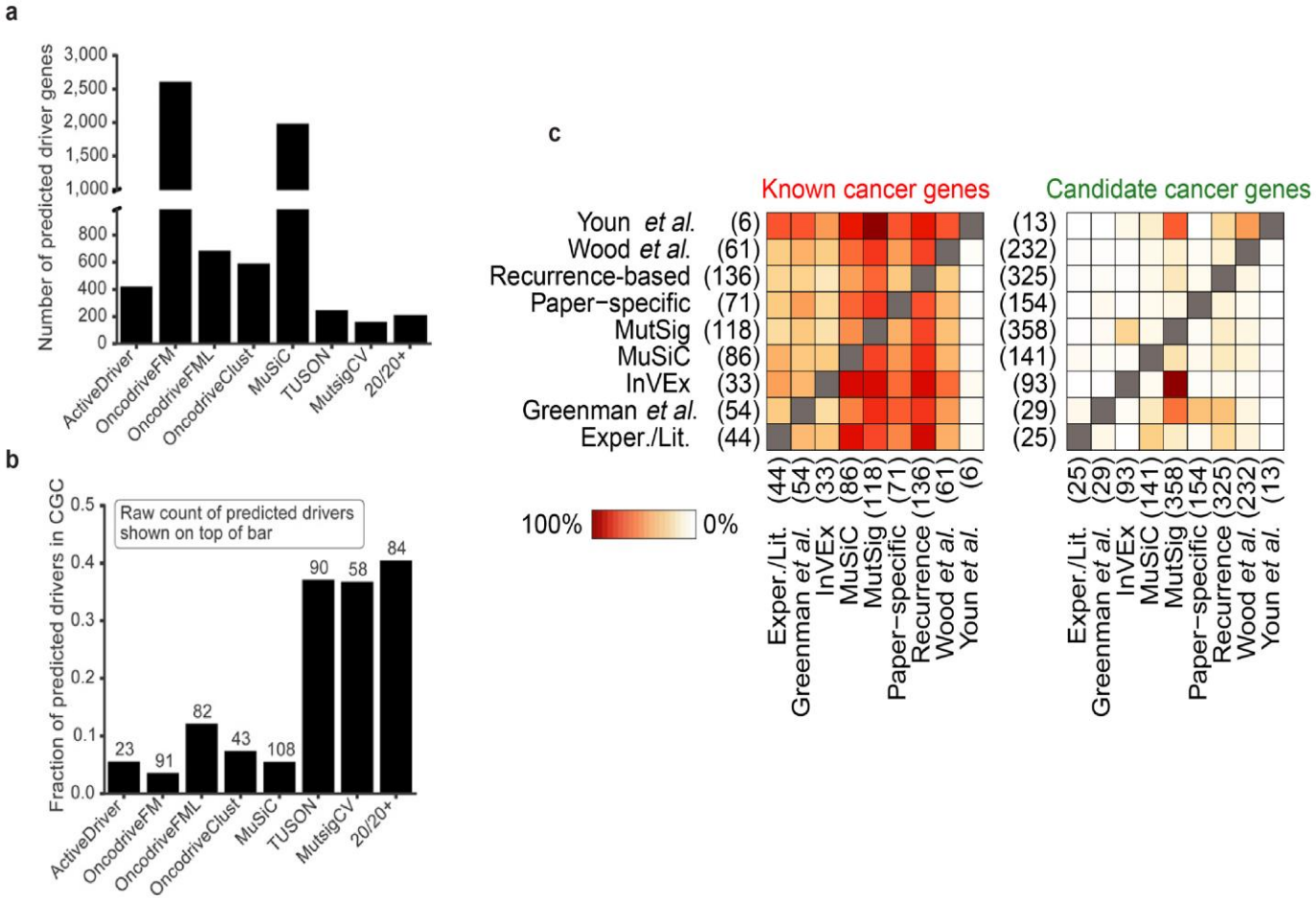

a
Number of predicted driver genes
3,000
2,500
2,000
1,500
1,000
800
600
400
200
0
ActiveDriver
OncodriveFM
OncodriveFML
OncodriveClust
MuSiC
TUSON
MutsigCV
20/20+
b
Fraction of predicted drivers in CGC
0.5
0.4
0.3
0.2
0.1
0.0
Raw count of predicted drivers shown on top of bar
23
91
82
43
108
90
58
84
ActiveDriver
OncodriveFM
OncodriveFML
OncodriveClust
MuSiC
TUSON
MutsigCV
20/20+
c
Known cancer genes
Candidate cancer genes
Youn et al. (6)
Wood et al. (61)
Recurrence-based (136)
Paper−specific (71)
MutSig (118)
MuSiC (86)
InVEx (33)
Greenman et al. (54)
Exper./Lit. (44)
(13)
(232)
(325)
(154)
(358)
(141)
(93)
(29)
(25)
100%
0%
Exper./Lit. (44)
Greenman et al. (54)
InVEx (33)
MuSiC (86)
MutSig (118)
Paper−specific (71)
Recurrence (136)
Wood et al. (61)
Youn et al. (6)
Exper./Lit. (25)
Greenman et al. (29)
InVEx (93)
MuSiC (141)
MutSig (358)
Paper−specific (154)
Recurrence (325)
Wood et al. (232)
Youn et al. (13)

Figure 1.2. Comparison of current methods for driver gene identification. **(a)** Outputs of the eight methods evaluated in Tokheim et al. (Tokheim et al. 2016). **(b)** Fraction of predicted driver genes ($q \leq 0.1$) for the same methods evaluated in a. The Cancer Gene Census (CGC) was downloaded April 1, 2016 (Forbes et al. 2015). Raw count of predicted driver genes are indicated on top of each bar. **(c)** Heatmaps showing the overlap between methods identifying known and candidate driver genes in the Network of Cancer Genes database (An et al. 2016). Each box represents the percentage of cancer genes identified with one method that are also supported by another. For each method, the total number of associated driver genes is reported in brackets. Figures a and b are taken from Tokheim et al. (Tokheim et al. 2016) and figure c is taken from An et al. (An et al. 2016).

1.3 Mutational landscapes and cancer drivers across tumour types

Unravelling the evolutionary history of tumours and, in particular, identifying the drivers, describing the order of their acquisition and understanding their functional relevance, have been the focus of a number of studies during the last decade (Garraway and Lander 2013; Kandoth et al. 2013; Lawrence et al. 2013; Vogelstein et al. 2013; Hiley et al. 2014). Dissection of cancer drivers is clinically relevant, as prognosis in some cases is linked to the mutational landscape, with certain types of alterations exhibiting better or worse prognosis than others. For example, in breast and oesophageal cancers amplifications of the oncogene *HER2* predicts for a good response to the anti-*HER2* monoclonal antibody trastuzumab (Bang et al. 2010; Romond et al. 2005). In contrast, extensive chromosomal instability has been associated to poor prognosis in multiple solid tumours (Carter et al. 2006; Walther, Houlston, and Tomlinson 2008) and metastasis (Bakhoum et al. 2018).

Despite the extraordinary efforts and the massive sequencing projects currently underway (International Cancer Genome Consortium et al. 2010), there is a number of fundamental questions that remain unanswered. One such question is the accurate estimation of the number of drivers required to convert a normal cell to a cancer cell and then sustain tumour progression (Renan

1993; Vogelstein et al. 2013). Initial estimates based on age-incidence statistics predicted that most cancers carry more than one drivers (Armitage and Doll 1954). In support of this, common adult epithelial cancers, such as colorectal and prostate, require five to seven rate-limiting events (*i.e.* drivers) to get established (Miller 1980; Stratton, Campbell, and Futreal 2009). The central premise of this approach is the assumption that all rate-limiting events are drivers, and consequently, all driver events are rate-limiting. Both assumptions were extensively challenged following evidence showing that selection is more important than increased mutation rate in terms of their ability to drive tumorigenesis (Tomlinson and Bodmer 1999; Tomlinson, Sasieni, and Bodmer 2002; Tomlinson, Novelli, and Bodmer 1996; Martincorena et al. 2017).

Moreover, it is still unclear whether there are any drivers exhibiting specificity for a particular cancer type. It has been demonstrated that the oncogenic fusion of BCR-ABL is strongly associated with the development of chronic myeloid leukaemia (Sawyers 1999). In contrast, lung adenocarcinoma shows a more diffused pattern of alterations in several signalling pathways, such as the receptor tyrosine kinase/RAS/RAF pathway, rather than recurrent alterations in specific genes (Cancer Genome Atlas Research Network 2014). Very recently Iranzo and colleagues (Iranzo, Martincorena, and Koonin 2017) examined the specificity of drivers using 7,665 samples, spanning 30 cancer types, and showed that drivers in several tumours, namely colorectal, pancreatic, endometrial, kidney (clear cell), breast, thyroid, and brain, were mostly tissue-specific. In contrast, other cancer types, including stomach, oesophagus and lung cancers, are characterised by a more diverse and less specific accumulation of drivers. This diffuse signal of drivers implies a high

inter-patient heterogeneity at the gene level and proves that the pattern of driver accumulation can be vastly different when comparing different cancer types.

Systematic analysis of pan-cancer cohorts revealed that different cancer types also exhibit different mutation frequencies (Kandoth et al. 2013; Lawrence et al. 2014; Vogelstein et al. 2013; Stratton, Campbell, and Futreal 2009). Mutagen-induced cancers, melanoma or lung cancer for example, tend to harbour many more mutations than genomically stable cancers, such as leukaemia, thyroid carcinoma, or paediatric tumours (Figure 1.3A). This heterogeneity is reflected at the functionally-relevant mutations as well, as damaging and truncating mutations are encountered in higher frequencies in mutagen-induced tumours (Figure 1.3B and C). In addition to the vast inter-tumour differences across cancer types, cancer samples exhibit significant heterogeneity even within the same tumour type. For instance, lung cancers associated with cigarette smoking harbour 10-fold higher mutations when compared to lung cancers in non-smoker individuals (Govindan et al. 2012). The number of somatic mutations present in tumour samples is tightly linked to the number of drivers. Although it would be expected that a high number of mutations would generate numerous drivers, recent studies suggested that the number of driver point mutations is surprising stable and low across cancer types, ranging from one to 10 with a pan-cancer average of seven (Sabarinathan et al. 2017). Although, these results are striking, further investigation is warranted to determine whether this phenomenon is an intrinsic property of cancer genomes or merely an underestimation of drivers due to limitations of the existing prediction algorithms.

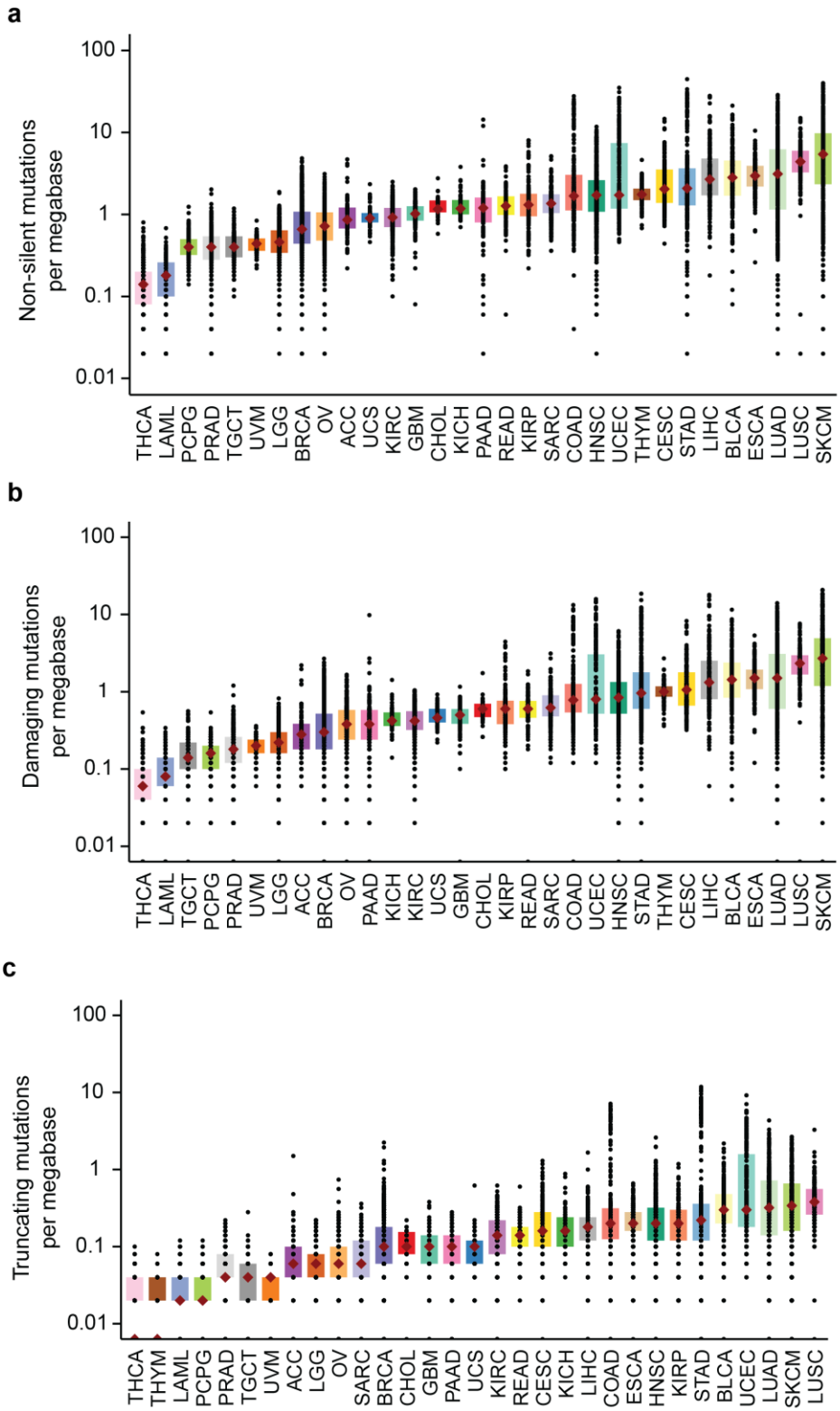


Figure 1.3. Distributions of **(a)** non-silent, **(b)** damaging, and **(c)** truncating mutations in 7,828 samples of 31 cancer types. In each plot, cancer types are sorted according to the median of the distribution (red diamond). Cancer type-specific colours are kept the same across all four plots to facilitate comparisons. Full names of cancer types and samples used are reported in Table 1.1.

Pan-cancer analysis of driver genes revealed a wide range of cellular processes that are typically deregulated in cancer (Kandoth et al. 2013). These processes can be broadly classified in 20 categories, shown in Figure 1.4, each of which is perturbed at various frequencies in individual tumour types. The most frequently mutated gene across all cancer types is *TP53,* which is altered in 42% of samples, with most mutations found in ovarian cancer (95%), lung squamous cell carcinoma (80%), head and neck squamous cell carcinoma (70%) and colorectal cancer (60%). *PIK3CA* is the second most frequently mutated gene, whose alterations are occurring in >10% of samples, with particular enrichment in uterine corpus endometrial carcinoma (50%) and breast cancer (30%). Many cancer types carry mutations in genes which are rarely found mutated in other cancer types. For instance, *FLT3* and *NPM1* are predominantly altered in acute myeloid leukaemia, *APC* and *KRAS* are primarily mutated in colorectal cancer and *VHL* is typically mutated in kidney renal clear cell carcinoma (Figure 1.4). It is worth pointing out that a wide range of genes belonging to the same category of processes, such as histone modifiers, are mutated in low or intermediate frequencies within the same cancer type. This highlights that recurrence of gene mutations is not the single most important criterion for driver identification and that selection may act in the level of deregulated pathways whose perturbations can in turn be used to identify processes that drive tumorigenesis.

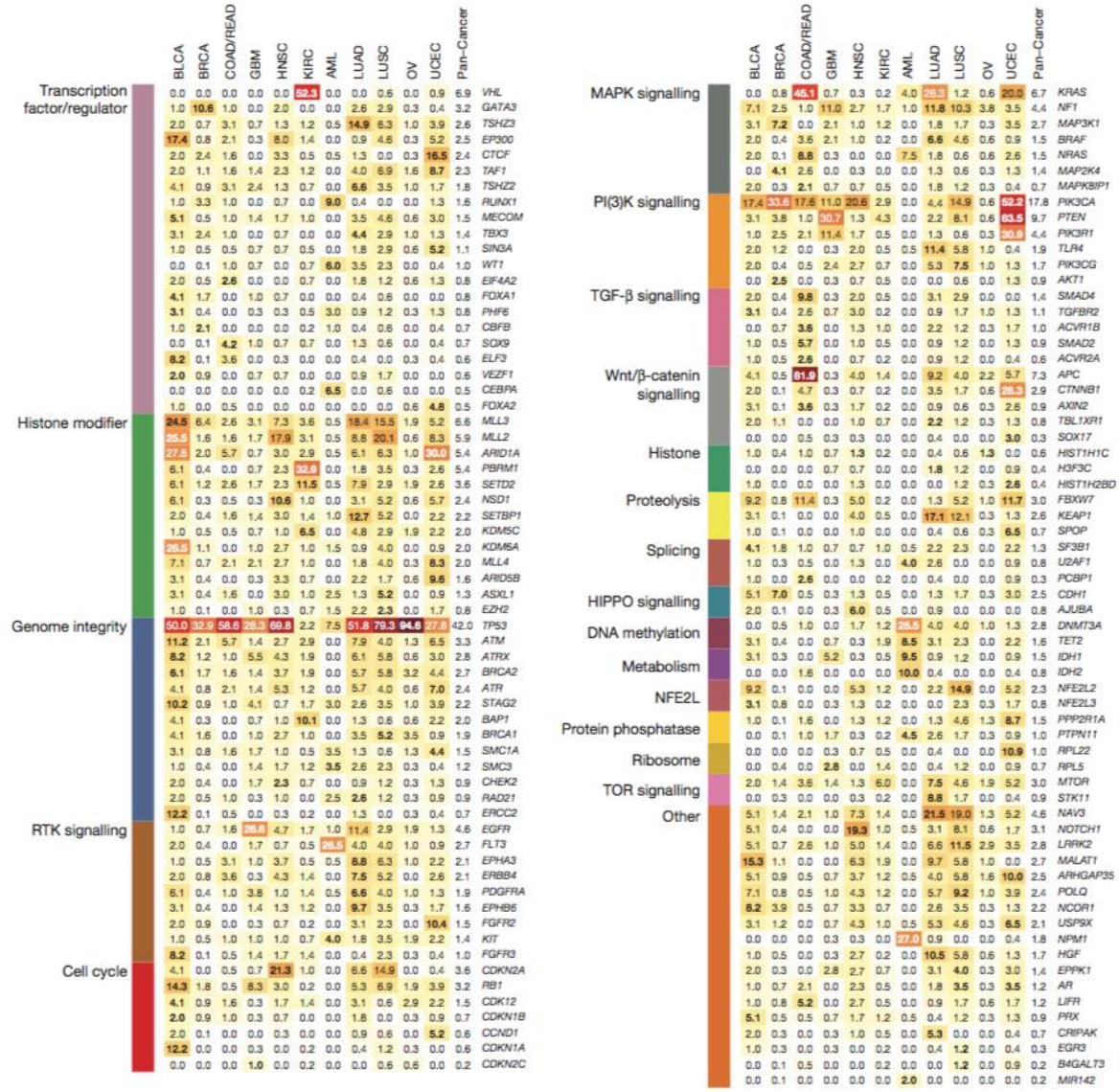


Figure 1.4. Pan-cancer identification of driver genes and their association with cellular processes. The 127 significantly mutated genes identified in 12 cancer types in Kandoth et al. (Kandoth et al. 2013) were broadly classified in 20 processes. The number in each cell represents the percentages of samples mutated in individual cancer types and pan-cancer. The highest percentage in each cancer type is in bold. Figure taken from Kandoth et al. (Kandoth et al. 2013). Full names of cancer types are reported in Table 1.1.

1.4 Systems-biology approaches to identify cancer drivers: towards personalised medicine and patient-specific driver genes

The functional heterogeneity of cancer drivers may initially suggest that each tumour is a distinct disease with partly unrelated genetic determinants.

However, this remarkable diversity of cancers does not necessarily imply absence of common properties or outcomes. It is well-documented that as normal cells progressively evolve to a neoplastic stage, they acquire alterations which affect “hallmark” capabilities. These hallmarks of cancer were described in 2000 by Hanahan and Weinberg and include sustaining of proliferative signalling, evasion of growth suppression, resistance to cell death, activation of invasion and metastasis, enabling of replicative immortality and induction of angiogenesis (Hanahan and Weinberg 2000). They were later updated to include emerging hallmarks, such as deregulation of cellular energetics and avoidance of immune destruction (Hanahan and Weinberg 2011). The description of these cancer hallmarks was ground-breaking as it was one of the first attempts to summarise the knowledge acquired from the inspection of the cancer genome and highlighted the fact that, irrespectively of individual mutations, there is a handful of key traits of cancer cells that encompass tumorigenesis.

Examining the driver potential of a cancer cell at a process level allows for a massive reduction of heterogeneity in individual gene level; from 20,000 human protein-coding genes to around two thousand pathways (Fabregat et al. 2016). This reduction is not only one of convenience, but it is based on the hypothesis that multiple single mutations can have the same outcome by altering different parts of the same biological process. For example, several cancer genes directly control the transition from G0 or G1 cell cycle phases to S phase, denoting a conversion from a resting cell stage to a replication stage (Vogelstein and Kinzler 2004). Another example that pinpoints the reasons why focusing on pathways instead of individual genes might reduce the heterogeneity when examining cancer genomes was provided by studies on

TP53. In many cancer types, it was shown that the most common way to disrupt *TP53* is by point mutations (Olivier, Hollstein, and Hainaut 2010). However, cancer cells can also inactivate *TP53* by amplification of *MDM2* or after infection by DNA tumour viruses (Vogelstein and Kinzler 2004). These intriguing examples of indirect inactivation of *TP53*-related processes generated interest in the interplay of the multiple components of biological pathways and the dissection of their relative contribution in cancer cells.

Apart from a simple description of pathways whose members are known driver genes, a systems-biology approach could assign a likely driver role to additional pathway components or even identify new pathways whose individual members are rarely mutated in cancer. Studies aiming to materialise on this hypothesis were assisted by the construction of comprehensive and accurate protein-protein interaction networks. Availability of high-throughput experimental data from complex organisms (Giot et al. 2003; Siming Li et al. 2004) enabled the first charting of interaction maps that consequently provided an insight into complex cellular functions and mechanisms (Jeong et al. 2001; Calvano et al. 2005). Furthermore, several computational methods were also developed with the objective to predict protein-protein interactions when experimental data were lacking or were insufficient to construct comprehensive interaction maps (Brown and Jurisica 2005; P. Jonsson et al. 2006).

Studies examined the properties of known driver genes in the context of human protein-protein interaction networks and found that these proteins are highly connected and occupy central positions in the interactome (Jonsson and Bates 2006; Rambaldi et al. 2008). These findings were interpreted as indicative of “fragility” of driver genes and triggered a wider search to identify global properties of driver genes. These properties designated “systems-level”

properties, are not strictly related to gene function in a cancer cell. They represent general attributes which render these genes different from the rest of human genes, and therefore, can be used to discriminate cancer drivers from passengers.

In addition to highly connected and central proteins, cancer genes are longer and they encode proteins with a higher number of protein domains on average (An et al. 2016; D'Antonio and Ciccarelli 2013). They tend to maintain a single copy in the genome (D'Antonio and Ciccarelli 2011; Rambaldi et al. 2008), and localise in heterochromatic regions of the genome. Moreover, evolutionary analysis showed that different types of cancer genes mostly appeared two times during evolution, with tumour suppressors originating in prokaryotes and oncogenes in metazoans (D'Antonio and Ciccarelli 2011; Rambaldi et al. 2008). This indicates that tumorigenesis arises from perturbation of either basic, as in the case of "old" tumour suppressor genes, or regulatory processes, as in the case of "young" oncogenes. Finally, cancer genes are regulated by a significantly higher number of miRNAs as compared to the rest of human genes (D'Antonio et al. 2012) and they tend to be ubiquitously expressed in normal human tissues (An et al. 2016; D'Antonio and Ciccarelli 2013).

Taken together, the properties of cancer genes described above, highlight a form of complexity that has not been taken into consideration when predicting cancer driver genes. The decipherment of this complexity can facilitate the transition from a gene-centric view of cancer to an integrated system of genes, proteins and processes that eventually all contribute to cancer via interactions and not as single entities.

1.5 Redefining cancer drivers: from a few major drivers to numerous “mini”-drivers

As cancer sequencing projects progressively incorporated higher number of samples, a striking feature of tumorigenesis became apparent: only a few new cancer driver genes have been found to be altered at high frequencies (Figure 1.1C). This observation highlighted the possibility that genes mutated in low number of samples, even in only one sample, may affect tumorigenesis and, in fact, they may act in addition to the few major drivers in cancer cells. This hypothesis was first described recently by Castro-Giner and colleagues (Castro-Giner, Ratcliffe, and Tomlinson 2015) and goes beyond the classical dichotomous description of genes mutated in cancer genomes as drivers and passengers. It provides an alternative perspective in which the driver potential of genes is continuous and includes a few major drivers and numerous genes that contribute in a modest way to tumour progression. These genes were designated “mini-drivers” as they represent an attenuated form of drivers and an amplified form of passengers.

A piece of evidence towards supporting this hypothesis was the finding that mutations, which previously were thought to be passengers, can be deleterious. In particular, McFarland and colleagues showed, using simulations and mathematical modelling of missense mutations, that moderately deleterious passenger mutations can be detected, and they have a predicted, but nevertheless, major effect on cancer progression (McFarland et al. 2013). This observation, even though it might have been connected to the overall passenger load of the cancer genome, implied that a number of passengers can play cumulatively a driver-like role. Although the examples of genes contributing to tumorigenesis as mini-drivers are limited, mainly due to the lack

of methods to identify them, there are examples of atypical mutations that are maintained in the cancer cell population and they seem to represent attenuated forms of known driver mutations. Examples of such mutations are changes in codons 146 and 117 of *KRAS* gene, which occur in a few cancer types (Smith et al. 2010). In contrast to typical driver mutations in *KRAS*, which occur in codons 12 and 13, these mutations are probably less effective in activating the RAS-RAF-MERK-ERK pathway (Smith et al. 2010; Castro-Giner, Ratcliffe, and Tomlinson 2015), and even when they co-occur, they are not driven out of cancer cell populations as a result of the presence of the main driver.

The roles of mini-drivers are possibly very broad, ranging from enhancing several hallmarks of cancer (Hanahan and Weinberg 2000, 2011) to fine-tuning the interplay of multiple drivers present in the same cancer cells. It is not inconceivable that “mini-drivers” could act by optimising the effect of major drivers, removing their negative side-effects or arise as a result of pressures present in the tumour microenvironment. Anecdotic examples for such functions exist in colorectal cancer, where the optimal level of WNT pathway is initially controlled by mutations in *APC* and is subsequently optimised by additional alterations (e.g. copy number variations) of *APC* (Segditsas et al. 2009) or inactivation of regulators of the pathways, such as *SOX9* (Castro-Giner, Ratcliffe, and Tomlinson 2015). Although the emergence of driver genes gives an overall selective advantage to the cancer cell, it also affects multiple pathways some of which may have contradicting functions. Mini-drivers may correct these contradicting functions and amplify tumorigenic signals. Finally, clonal expansions of tumour cells result in a highly heterogeneous population of cells some of which are expected to carry mini-drivers. Therefore, it is likely that

there could be a co-operative relationship between cancer cells with only drivers and cancer cells with drivers and mini-drivers.

There are several lines of evidence to support the mini-driver hypothesis (Bennett et al. 2018). However, examples are fairly limited to modifications of functions or specific mutations of known cancer genes. This can be explained by the fact that the ability to detect genes, which are mutated in such low frequencies, if not in single samples, is controlled by the sample size. It is, therefore, of interest to develop new methods to discover rare drivers or mini-drivers, especially in cancer types in which known drivers are very few and present in a relatively small fraction of samples. Detection of mini-drivers will allow us to prove or disprove another hypothesis, under which multiple weak driver events can substitute a major driver. This is of particular importance to certain tumours for which no major driver effect is predicted using the currently existing methods (approximately 10% of all cancer samples).

1.6 Aims of the thesis

The main aim of this thesis was to develop a computational algorithm to detect cancer driver genes in individual samples. Computational methods developed to date are focused on the identification of genes that are subject to positive selection and are recurrently altered across cancer samples. In contrast, the algorithm presented in this thesis operates within each cancer sample and therefore, allows the discovery of patient-specific or rare cancer drivers.

Extending the classical view of distinction between drivers and passengers, the working hypothesis of this thesis was that the cancer driver

potential is, in fact, continuous with few major drivers and numerous genes contributing to different degrees to tumorigenesis. To identify novel and patient-specific cancer drivers, I employed supervised machine learning and, in particular, support vector machines. The development phase of my algorithm is documented in the second chapter of this thesis, in which I describe in detail my efforts to characterise the distinctive features of known drivers that are subsequently used to train a one-class support vector classifier. In the third chapter, I describe the application of my algorithm to a cohort of 261 oesophageal adenocarcinomas (OACs) from the International Cancer Genome Consortium (ICGC). This allowed the refinement of the algorithm and the identification of its limitations and future extensions. Finally, in the fourth chapter of this thesis, I describe the in-depth characterisation and experimental validation of the predicted drivers and their utility in patient stratification and therapeutic intervention in OAC.

Chapter 2. Development of one-class systems-level support vector machine to predict cancer drivers in individual patients

2.1 Chapter overview

In this chapter I describe the development of a novel algorithm that predicts patient-specific cancer driver genes. The algorithm, which was named sysSVM because it utilises systems-level properties of human genes and support vector machines, was developed using a pilot cohort of 18 OACs from ICGC. After introducing the mathematical framework and the features of support vector machines, I describe the algorithm itself, the preliminary predictions in the pilot cohort and their relevance for OAC, as well as its comparison with other existing algorithms.

2.2 Introduction

The increasing amount of data produced by high-throughput sequencing and imaging platforms require the development of new computational approaches for the exploration of the complex landscape of biological systems. Integrations of data across platforms can be facilitated by machine learning algorithms, which are capable of extracting high-dimensional patterns (Li, Wu, and Ngom 2016). These patterns are rules that describe the data in the form of mathematical functions. Machine learning systems can learn these patterns on their own by close examination of large datasets, where the patterns of interest are present. The process of learning begins with a set of observations, which are supplied as examples and comprise the training set. These examples are

then used to make decisions based on their similarity to unseen data. As a data-driven approach, machine learning is dependent on large and well-structured trainings sets to be able to predict outcomes with confidence.

Machine learning is a mathematical modelling process that uses equations (models) with carefully selected terms (parameters) to make predictions or calculate probabilities of possible outcomes, based on a n-dimensional vector of variables (features) that represent the training set. These parameters are selected *via* a trial-and-error process (optimisation), during which their values are selected based on the features of the training set.

Different machine learning algorithms have been successfully used in biological research to discover biomarkers and subgroups of patients from gene expression data (Ramaswamy et al. 2001; Khan et al. 2001; Tibshirani et al. 2002; Golub et al. 1999; Pal et al. 2007). Additionally, they have recently been applied in translational research to extract information from imaging technologies and improve diagnostic and prognostic accuracy (Gillies, Kinahan, and Hricak 2016). In general, machine learning algorithms can be used to estimate a continuous value, for example the probability of a patient to relapse, or for classification purposes, i.e. to assign observations to pre-defined categories.

Machine learning algorithms can be broadly divided in unsupervised, semi-supervised and supervised. Unsupervised methods are used to detect patterns in the data when no ground truth or known observations related to these patterns are available (Hastie, Tibshirani, and Friedman 2009). Unsupervised tasks can be further categorised into two groups: i) clustering, in which the inherent grouping of the data is of interest, and ii) association problems, in which rules describing large portions of the data are derived. Many

popular clustering algorithms belong to unsupervised methods, with most prominent examples being hierarchical and k-means clustering (Rokach and Maimon 2005).

Semi-supervised algorithms are applicable when data of known examples are limited and disproportionally less than the unknown observations. Many real-world applications of machine learning belong to this category, as it is usually time consuming and expensive to label data using expert knowledge (Chawla and Karakoulas 2005; Blum and Mitchell 1998).

Finally, the majority of applications of machine learning use supervised learning, in which mathematical models are trained using labelled observations and they are subsequently used to categorise unlabelled data (Kotsiantis 2007). In many classification problems pre-defined rules do not exist and therefore, rules should be derived by a set of observations, which are considered to be representative of the objects that will be consequently classified. In the simplest scenario, there are two opposing classes of observations that are used during training, true positives and true negatives, but it is also possible to build supervised classifiers that derive rules from? additional classes of observations (Tsoumakas, Tsoumakas, and Katakis 2007). Often the type of function used to build a classifier is chosen beforehand and its parameters are optimised during training. Examples of these functions are linear/logistic regression (Freedman 2009), neural networks (Hopfield 1982) and support vector machines (Cortes and Vapnik 1995). The optimisation phase is usually an iterative process, during which multiple subsets of the labelled data are used and the robustness of various parameters across different sets of training observations is assessed.

The choice of the most suitable classification algorithm is not a trivial issue, as different algorithms have distinctive strengths and weaknesses and

therefore, no single algorithm works best for all classification problems. The (usually limited) set of training observations might not resemble in full the true positive and negative observations in real life. In this case, the distributions of their features are not sufficiently separable in order to extract classification rules with high confidence. This issue is of particular importance for data derived from biological experiments, which are usually limited and noisy. Consequently, the optimal classifier has to be flexible enough to fit the data, but also robust enough to exclude random and uncharacteristic noise (Geman, Bienenstock, and Doursat 1992).

A well-known issue in the selection of classification algorithms is the bias-variance dilemma (Geman, Bienenstock, and Doursat 1992). This states that there is a trade-off between the bias and variance of a classification algorithm that needs to be considered and optimised during training. In a case where there are multiple training sets for a classification problem, which are equally good, variance refers to the amount by which the optimal function of the classifier will change when different training sets are used. Hence, variance reflects the adaptability of the algorithm to different training sets. On the other hand, bias refers to the error that is introduced by the algorithm during prediction and reflects its ability to model real life (Brighton and Gigerenzer 2015). A flexible classification algorithm is highly adaptable and therefore, has low bias, but, if it is too flexible, it will model the training observations too closely. This problem is known as overfitting. Overfitted classification functions achieve high performance in very specific training sets, but their generalisation performance is very poor. The issue of overfitting is also affected by the number of features and the number of observations in the training set. Small training sets with high number of features are prone to overfit, a phenomenon which is

known as the curse of dimensionality (Cabestany et al. 2005). The most common solution to address the curse of dimensionality is feature selection, a process which retains only the most informative features for a given training set (Pudil, Novovieova, and Kittler 1994; Jain and Zongker 1997).

Overall, although various classification algorithms and optimisation techniques are available, it is often difficult to derive classification rules that describe the training observations completely. Even the best classification algorithm for a given problem might leave regions of the feature space inadequately described. The inclusion of additional (and possibly suboptimal) classifiers can often help by capturing valuable information that is neglected by the single best classifier (Kuncheva 2004). In the rest of this chapter, after a short introduction of support vector machines and one-class classifiers, I describe the development of a meta-classifier, named sysSVM, which combines four classifiers using four different mathematical functions to predict cancer driver genes in individual patients.

2.2.1 Support Vector Machines

Support Vector Machines (SVMs) refer to a family of supervised machine learning algorithms that since their initial development in the 1990s (Cortes and Vapnik 1995) have been used for classification in a variety of tasks (Wang et al. 2010; Chen et al. 2009; Li et al. 2003). SVMs try to address and generalise the problem of separating observations using a decision boundary in a high-dimensional space (James 2013) (Figure 2.1A). The separation boundary can then be used to classify unknown observations (Figure 2.1B). In their simplest version, that of one-dimensional space, the separation hyperplane is a single

point (Figure 2.1C). In two dimensions it is a line and, in general, in a p-dimensional space, the separation hyperplane is a subspace of p-1 dimensions (hyperplane; Figure 2.1D) and its mathematical definition follows the equation:

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 = 0 \quad (2.1)$$

with parameters β_0, β_1 and β_2. Equation 2.1 describes a line and it can be extended to p-dimensional problems by adding the relevant parameters as follows:

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_p X_p = 0 \quad (2.2)$$

For a given point $X = (X_1, X_2, X_3, \ldots, X_p)^T$, if X satisfies 2.2 then X lies on the separation hyperplane. It is therefore easily conceivable that observations that are located in either side of the hyperplane will satisfy one of the following two conditions:

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_p X_p > 0 \quad (2.3)$$

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_p X_p < 0 \quad (2.4)$$

and therefore, the side in which each observation lies can be determined by calculating the sign of 2.2.

If such a hyperplane exists, classification problems can be solved in a very intuitive manner by assigning two classes, for example +1 and -1 (see

equations 2.5 and 2.6 below), to observations with a positive or negative sign of equation 2.2, respectively (Schölkopf and Smola 2002). In fact, there will be an infinite number of hyperplanes that separate a given set of data (Figure 2.1E). This is because minor shifts of a hyperplane, where shifts can refer to rotations or small displacements, can lead to new hyperplanes. Therefore, selection criteria should be established in order to select the optimal hyperplane. A natural choice would be the hyperplane that is equally far from all classes of observations (i.e. positive and negative), and this is the maximal margin hyperplane (Figure 2.1F). Margin (M) refers to the minimum distance of the hyperplane from the training observations and one of the main objectives during training and parameter optimisation is to maximise M such that:

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_p X_p > 0 \text{ if } y_i = +1 \quad (2.5)$$

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_p X_p < 0 \text{ if } y_i = -1 \quad (2.6)$$

$$y_i(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \cdots + \beta_p X_{ip}) \geq M \quad \forall i = 1, \ldots, n \quad (2.7)$$

This optimisation problem in equation 2.7 defines the constraint that guarantees that all observations reside in the correct side of the hyperplane with equal or greater distance than M.

In many cases, it is desirable to deviate from such a strict definition of the constraint in equation 2.7 to avoid overfitting the hyperplane. An overfitted hyperplane describes too closely a given set of data and lacks the ability to be generalised and accurately predict future data (James 2013). A relaxed constraint would allow some of the observations in the training set to reside in

the wrong side of the hyperplane, accounting for natural variability in the data and providing robustness to individual observations (Figure 2.1G). This is the soft margin hyperplane (Figure 2.1H) and its optimisation is described by the following equations:

$$y_i(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p) \geq M\,(1 - e_1)\ \forall i = 1, \ldots, n \qquad (2.7)$$

$$e_1 \geq 0, \sum_{i=0}^{n} e_1 \leq C \qquad (2.8)$$

where *e* allows for training observations to be on the wrong side of the hyperplane and *C* is a non-negative tuning parameter (Chapelle and Bousquet 2002). The role of parameter *C* is to control the overall severity of the constraint in equation 2.7, as it bounds the sum of *e*. It is optimised during cross-validation where the training set is split into training and test sets and multiple values of *C* are tested as upper bounds of 2.8. Small values of *C* (e.g. close to zero) define narrow margins that are rarely violated by training observations and, therefore, fit the given dataset closely. The higher the value of *C*, the softer the margin, thereby allowing for higher values of *e* (2.8) and more training observations to reside on the wrong side of the separation boundary. The constraint in equation 2.7, owing to the presence of *e*, has the property of being sensitive only to the observations that are either located in the margin or those that violate the hyperplane. These observations are called support vectors. This is a unique feature of SVMs as observations that are on the correct side of the separation hyperplane (and far from the margin) do not affect the formation of the hyperplane.

The boundaries described in the introduction of this chapter are all linear but, in many cases when biological data are concerned, a linear separation hyperplane does not exist (Figure 2.1I). These cases are often referred to as non-separable and a class of non-linear (or very soft linear) boundaries are employed instead (Suykens 2001). When the observations of interest are non-separable, the constraint in equation 2.7 has no solution with $M > 0$. In these cases, SVMs employ helper functions (kernels) (Schölkopf and Smola 2002) in order to enlarge the feature space and solve 2.7 (Figure 2.1J-L). For instance, instead of trying to separate a given set of observations in a p-dimensional space, one could add quadratic versions of the features. This enlargement of the feature space might make positive and negative observations separable. There are multiple kernels of particular interest in this thesis that are widely used in literature. As explained in detail below, four kernels are used in sysSVM, namely linear, polynomial, radial and sigmoid. These four kernels are combined in a meta-classifier, in an effort to alleviate misclassifications by individual kernels and to prioritise genes that are predicted as positive by multiple kernels.

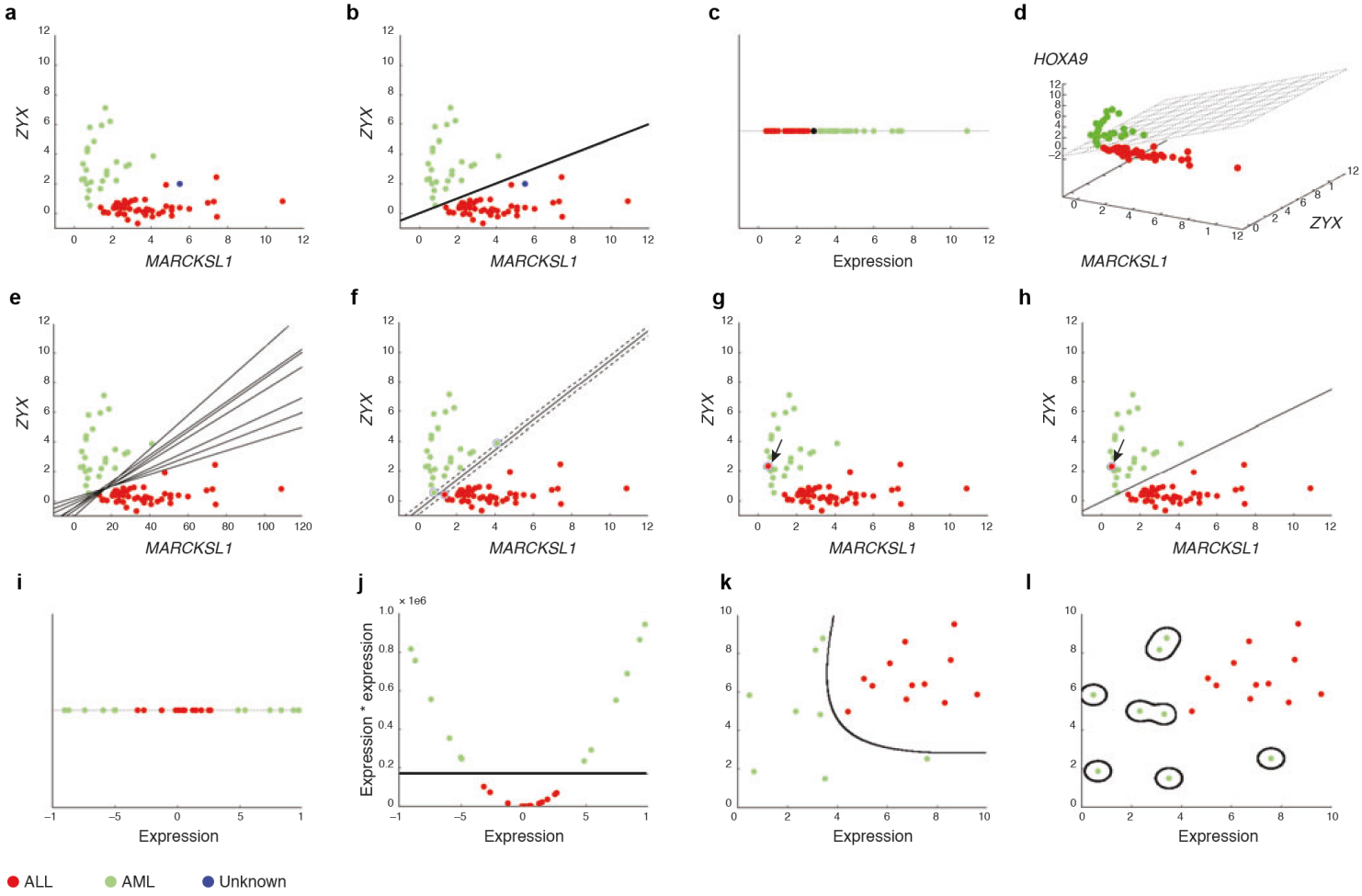

a
b
c
d
e
f
g
h
i
j
k
l
ZYX
MARCKSL1
Expression
HOXA9
Expression * expression
× 1e6
ALL
AML
Unknown

Figure 2.1. An illustration of the main characteristics of support vector machines (SVM). Figure taken from (Noble 2006). **(a)** mRNA expression level of ZYX and MARCKSL1 in lymphoblastic leukaemia (ALL; red) and acute myeloid leukaemia (AML; green) patients. The main task of SVM is to assign a label (ALL or AML) to the unknown observation (blue). **(b)** A separating boundary (hyperplane) is drawn between ALL and AML. Based on this hyperplane the label of the unknown observation is “ALL”. **(c)** A hyperplane in one dimension. It is shown as a black point. **(d)** A hyperplane in three dimensions. **(e)** Many possible hyperplanes in two dimensions. **(f)** The maximum-margin hyperplane. The support vectors are circled. **(g)** Same data set as in a with one additional ALL patient. This patient is considered error or outlier because the expression profile resembles that of AML. **(h)** A soft margin separation hyperplane. **(i)** A non-separable one-dimension data set. **(j)** Separating data in i by applying a kernel that adds another variable, the square of expression (y axis) in the data set. **(k)** A non-linear separation hyperplane. **(l)** A non-linear separation hyperplane which potentially overfits the data.

2.2.2 One-class classification

Although in many cases classification problems can be described as binary (or multi-categorical) problems of positive and negative observations, there are cases for which only one class of observations is available. In such cases, the defined problem is the description of the target set of observations and the detection of all new observations that resemble the target set. Alternatively, the problem can also be defined as identifying all observations that are not similar with the target set. In a one-class classification framework both cases are identical.

In literature, one-class classification has been described using numerous terms. The term “one-class classification” was first described in Moya et al. (Moya, Koch, and Hostetler 1993) but later, owing to the many applications of the method, it was also described as “outlier detection” (Ritter and Gallegos 1997) and “novelty detection” (Bishop 1994). The first application of one-class classification was to optimise the decision boundaries and enhance out-of-class

generalisation of neural networks, with particular emphasis on minimising the size of the decision boundary in order to produce small classification mapping errors (Moya, Koch, and Hostetler 1993). As classifiers provide reliable estimates for unknown observations resembling the training set, extrapolations to sparse regions of the feature space are of low quality (Roberts, Penny, and Pillot 1996). In such cases, outlier detection and elimination should first be used before classification and prediction.

Based on their statistical foundation, one-class algorithms can be broadly divided in i) proximity-based, ii) parametric methods, and iii) semi-parametric methods (Hodge and Austin 2004). Proximity-based techniques, such as k-Nearest Neighbour (k-NN), are easier to implement, but their computational complexity increases proportionally to the dimensions of the data. Therefore, they need extensive optimisation. Parametric methods model the density of the data in the feature space and their complexity grows with the complexity of the model, rather than the size of the data. A well-known algorithm in this category is the Minimum Volume Ellipsoid estimation (MVE) (Rousseeuw 1985), which calculates the smallest ellipsoid volume around the majority of the data distribution model, representing the high-density areas of the feature space. Finally, semi-parametric methods apply local kernel models, instead of a single distribution model, to the whole data set and identify the outliers as observations located in regions of low density (Hodge and Austin 2004). One-class support vector machines belong to this last category as they project the input data onto high dimensional spaces using a variety of kernel functions.

2.3 Algorithm development

2.3.1 Selection of predictive features of known driver genes

In chapter 1, I introduced numerous features of known driver genes that are collectively referred to as systems-level properties (D'Antonio and Ciccarelli 2013; An et al. 2016; Rambaldi et al. 2008; D'Antonio and Ciccarelli 2011; Domazet-Lošo and Tautz 2010; Jonsson et al. 2006). The objective of this part of my thesis was to survey which of the systems-level properties can serve as predictors of novel driver genes and subsequently use them to develop a c-lassifier. To this end, I hypothesised that the best features would be those that exhibit statistically significant difference between known driver genes and the rest of human genes.

Starting from a total of 19,014 human protein-coding genes (see methods), I categorised 518 genes as known drivers according to the Cancer Gene Census (Forbes et al. 2017) and the remaining 18,496 comprised the rest of human genes. For each group, I annotated the corresponding system-level properties, as previously described (An et al. 2016), and broadly classified them in one of the following seven categories:

i) Gene duplicability and evolutionary origin: number of copies that a gene maintained in the genome and the node of the tree of life that the oldest ortholog of a gene has been found in. Additionally, I collected data on whole genome duplication events from recent literature (Makino, McLysaght, and Kawata 2013)

ii) Gene and protein expression: the number of human tissues that a gene was expressed in according to the Genotype-Tissue Expression database (GTEx) (Lonsdale et al. 2013)

iii) Protein-protein and miRNA-target interactions: properties that I derived from the human protein-protein interaction network, such as degree and betweenness, and the number of miRNAs interacting with a gene (Keshava Prasad et al. 2009; Ruepp et al. 2009; Milacic et al. 2012). I also assessed the membership of each gene to protein complexes

iv) Number of protein domains: protein domains in each gene as reported in the Conserved Domains Database (CDD) database (Marchler-Bauer et al. 2011)

v) Chromatin accessibility: data for the chromatin compartment that each gene was located in from Mutation Significance database (MutSigDB; https://software.broadinstitute.org/gsea/msigdb/)

vi) Replication timing: cell-cycle phase (e.g. early, intermediate or late) at which a gene was replicated

vii) Gene length (D'Antonio and Ciccarelli 2013)

Recent reports suggested that the chromatin compartment, in which a gene is located, and its replication timing during the cell cycle contribute to cancer mutational landscapes, due to restrictions to the accessibility of repairing enzymes and the depletion of the nucleotide pool of the cell, respectively (Lawrence et al. 2013). These properties have been previously implicated in algorithms for cancer driver discovery (Lawrence et al. 2013; Tokheim et al. 2016).

I next assessed the missing data for each property, and I excluded properties with no available data for a significant fraction of human genes from downstream analysis. I found that the systems-level property with the highest number of missing data was the membership in protein complexes, which was

not available for 70% of human genes (Figure 2.2A). The majority of human genes had incomplete data in one or two systems-level properties (Figure 2.2B), while only a very small proportion had missing data in more than two properties. To account for this, I performed median imputation for continuous properties and mode imputation for the categorical ones (Enders 2010). Specifically, for each property median or mode values were calculated for known drivers and the rest of human genes. All missing values were replaced for each gene with the corresponding value of the gene group in which it belonged to.

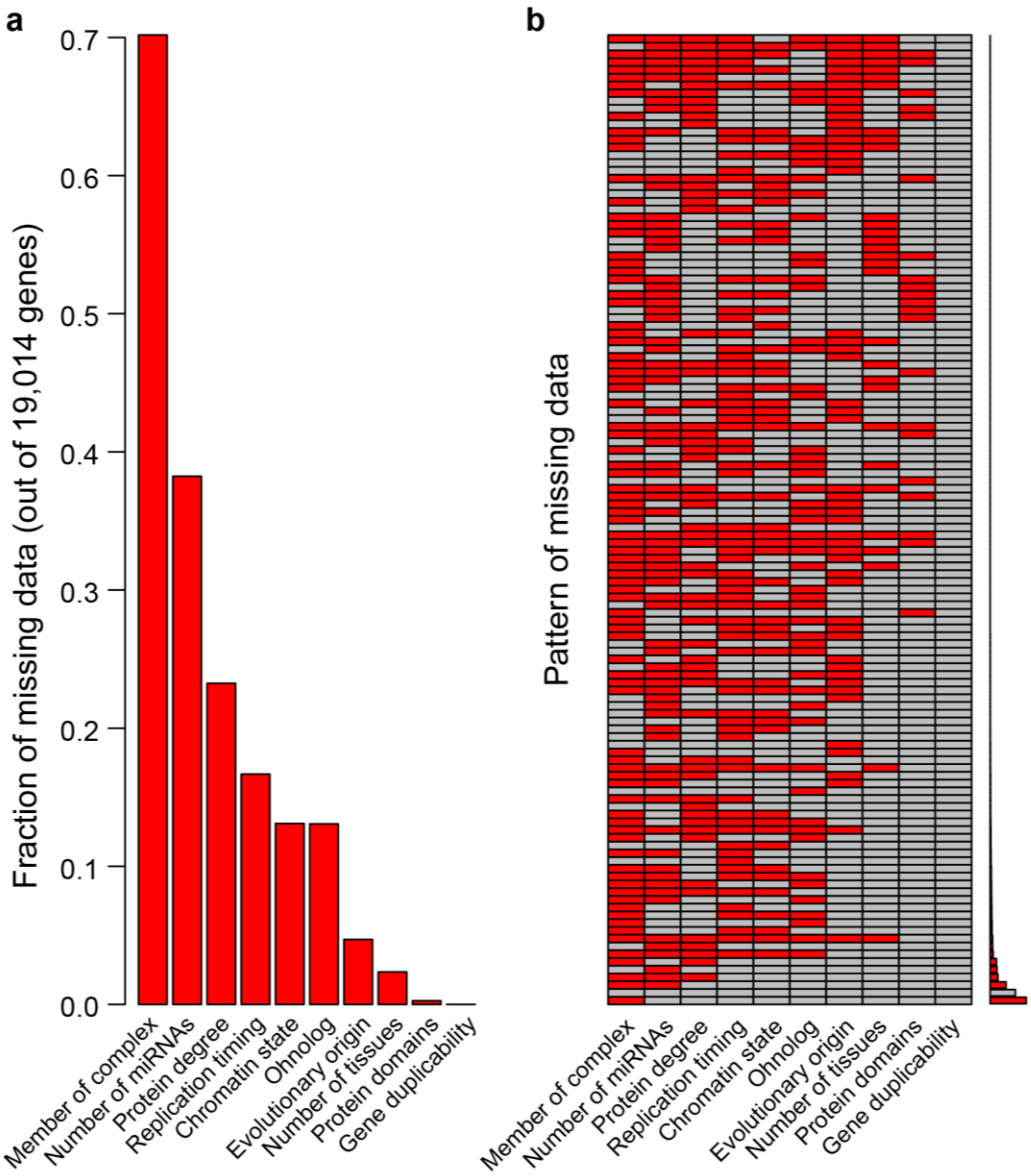


Figure 2.2. Missing data in systems-level properties of human genes as percentage **(a)** and pattern of combinations of properties **(b)**.

I separately tested the distribution of each property in known drivers against the rest of human genes and I only considered properties with significant difference between the two groups for the development of the classifier. In accordance to previous studies (An et al. 2016), these properties exhibited significant difference between the two groups of genes with the only

exception being the replication time, which was excluded from downstream analysis (Figure 2.3).

During the development of sysSVM, I decomposed systems-level properties to multiple features and included several of them in the final classifier, as both continuous and binary features. For example, I converted interactions to 4 features, namely degree, betweenness, hub and central protein. Hubs were genes whose proteins belonged to the top 25% of the most connected proteins in the human protein-protein interaction network as derived from the degree distribution. Similarly, I considered as central proteins those in the top 25th percentile of the betweenness distribution. Although, the continuous and the binary variants of the same feature carried the same type of information for the classifier, I reasoned that the binary features supplied a qualitative threshold of the distributions of the corresponding continuous variables, which could not be accurately captured when only continuous variables were used. Additionally, I converted single multi-level categorical properties to multiple binary ones. The final set of systems-level features was comprised of 24 properties (Table 2.1). These were complemented by molecular features that I derived from sequencing data of individual cancer samples. In particular, these molecular features included somatic alterations with predicted damaging effect on the protein function, i.e. truncating and non-truncating damaging mutations, gain-of-function mutations, gene gains and losses (see Methods). In total, 7 molecular features were additionally included in sysSVM (Table 2.1 and Methods), increasing the number of features to a total of 31.

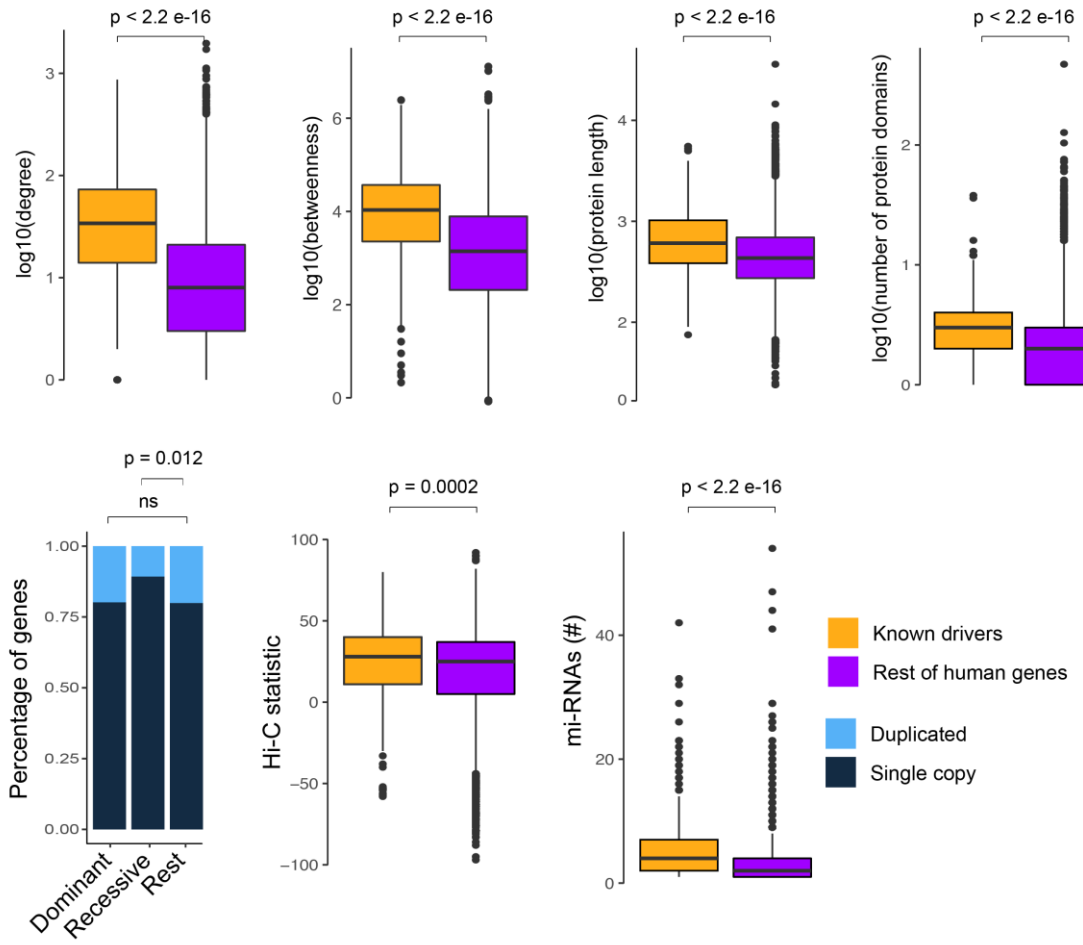


Figure 2.3. Systems-level properties. Distribution of degree and betweenness in the human protein-protein interaction network, the protein length, the number of protein domains, the percentage of genes maintaining a single or multiple copies in the genome, the chromatin compartment (measured from Hi-C experiment) and the number of mi-RNAs regulating the function of each gene were compared between known cancer drivers and the rest of human genes. For the comparison of gene copies, genes were also divided by mode of inheritance. Known cancer driver genes (n=505) were retrieved from the Cancer Gene Census (Tate et al. 2018) and their associated systems-level properties from the Network of Cancer Genes database (An et al. 2016). These properties were also annotated for the rest of human genes. P-values were estimated using a Wilcoxon Rank Sum test for all comparisons except for the duplication status for which Fisher's exact test was used.

Table 2.1. Summary of features used during the development of sysSVM. For each feature, the type and description are reported. In total, 31 features were used in sysSVM to predict cancer genes in the pilot cohort.

Features	N	Type	Description
Gene length	1.	Systems-level (continuous)	Length of the coding sequence (longest RefSeq)
Gene duplicability	2.	Systems-level (categorical)	≥1 gene duplicate ≥60% protein coverage
Ohnolog	3.	Systems-level (categorical)	≥1 duplicate from whole genome duplication
Number of protein domains	4.	Systems-level (continuous)	Total number of CDD domains (longest RefSeq)
Protein degree	5.	Systems-level (continuous)	Number of connections in the PIN
Protein betweenness	6.	Systems-level (continuous)	Centrality in the PIN
Hub	7.	Systems-level (categorical)	Top 25% most connected proteins (≥23)
Central protein	8.	Systems-level (categorical)	Top 25% most central proteins (≥6105)
Old gene	9.	Systems-level (categorical)	Origin before metazoans
Origin in prokaryotes	10.	Systems-level (categorical)	Oldest ortholog found in prokaryotes
Origin in single cell eukaryotes	11.	Systems-level (categorical)	Oldest ortholog found in eukaryotes
Origin in opisthokonts	12.	Systems-level (categorical)	Oldest ortholog found in opisthokonts
Origin in metazoans	13.	Systems-level (categorical)	Oldest ortholog found in metazoans
Origin in vertebrates	14.	Systems-level (categorical)	Oldest ortholog found in vertebrates
Origin in mammals	15.	Systems-level (categorical)	Oldest ortholog found in mammals
Origin in primates	16.	Systems-level (categorical)	Oldest ortholog found in primates
Ubiquitously expressed	17.	Systems-level (categorical)	Expressed in > 28/30 tissues (GTEx 4)
Medium expressed	18.	Systems-level (categorical)	Expressed in 3-28 tissues (GTEx 4)
Selectively expressed	19.	Systems-level (categorical)	Expressed in 2-3 tissues (GTEx 4)
Specifically expressed	20.	Systems-level (categorical)	Expressed in 1/30 tissues (GTEx 4)
Not expressed	21.	Systems-level (categorical)	Expressed in 0/30 tissues (GTEx 4)
Number of tissues	22.	Systems-level (continuous)	Number of tissues where the gene is expressed (GTEx 4)
Number of miRNAs	23.	Systems-level (continuous)	Number of miRNA interactions

Chromatin state	24.	Systems-level (continuous)	HiC statistic (Lieberman-Aiden Science 2009)
Exonic SNVs and Indels	1.	Molecular (continuous)	Silent and non-silent mutations (ANNOVAR)
Truncating mutations	2.	Molecular (continuous)	Stopgain, stoploss, frameshift alterations (ANNOVAR)
Non-truncating damaging mutations	3.	Molecular (continuous)	Damaging non-frameshift, nonsynonymous, splicing alterations (dbNSFP)
Gain of function mutations	4.	Molecular (continuous)	Gain of function (OncoDriveClust)
Gene copy number	5.	Molecular (continuous)	Segment mean from ASCAT
Gene loss	6.	Molecular (categorical)	Copy number 0 or 1
Gene gain	7.	Molecular (categorical)	Copy number >=4
Total	**31**	**-**	**-**

2.3.2 Description of the pilot sample cohort

As a pilot study to develop the classifier, I used a cohort of 18 OACs from ICGC (hereafter referred to as pilot cohort) for which somatic mutations and copy number variation data were available. The low number of known drivers in oesophageal adenocarcinoma in combination with the disappointing results in most recent clinical trials rendered this cancer type a particularly pertinent sample cohort for the development of the sysSVM method. However, the aim was to develop a method which was not cancer type-specific, but easily applicable to multiple cancer types. As it will become apparent in the following paragraphs, the only cancer type-specific part of my method was the construction of the training set, and by altering it, the method can be applied to any cancer type.

The pilot cohort was comprised of 16 male and two female patients, 12 of which had no prior treatment, while six were of unknown treatment status. Fourteen of the specimens (77.8%) were of distal oesophagus (siewert type 1 & 2) with the remaining four being either of subcardial stomach (siewert type 3) or of unknown origin. The average age at diagnosis was 69.7 years. One patient was diagnosed with distant metastasis and seven patients were diagnosed with metastasis to regional lymph nodes. A summary of all the clinical characteristics of the pilot cohort is reported in Table 2.2.

Table 2.2. Summary of the clinical characteristics of the pilot cohort (18 OAC samples). Percentages that do not sum up to 100% within each category denote missing data. Data were collected from the ICGC. Gastro-oesophageal junction (GOJ) types are reported according to Siewert classification.

Mean age at diagnosis (stDev)	69.7 (10.25)
Sex (% male)	88.8
Treatment (%)	
No treatment	66.7
Unknown	33.3
Tumour location (%)	
GOJ Type 1	50
GOJ Type 2	27.8
GOJ Type 3	11.1
Primary tumour (%)	
Stage I	27.8
Stage II	11.1
Stage III	38.9
Stage IV	0.6
Node positive (%)	44.4
Metastasis (%)	0.6

2.3.3 Systems-level one-class support vector machine (sysSVM)

Classical approaches for the development of supervised classifiers have relied on the presence of both positive and negative observations during model training. In this case this was not feasible, as negative observations, i.e. a set of confirmed non-driver (passenger) genes, could not be defined. In fact, even if a set of negative observations existed (see below), it would not have been representative of all passenger human genes, which was the desirable attribute

of the training set for model training. The discovery of genes exhibiting such patterns was one of the main aims of this thesis, as mentioned previously. Of note, a handful of possible non-driver genes have been described in recent literature, mainly showing that their function was irrelevant to tumorigenesis, such as in the case of olfactory receptors (Lawrence et al. 2013). The usage of these genes as negative observations would have led to an erroneous decision boundary, as their properties would not have been representative of those of all passenger genes. As a result, many genes whose properties would have not resemble those of the known cancer genes would have been predicted as cancer genes by the classifier, because the decision boundary was built to discriminate against the few genes in the non-representative negative set (Figure 2.4). This phenomenon was observed in the initial stages of my project, during which I started formulating my algorithm as a two-class classification problem.

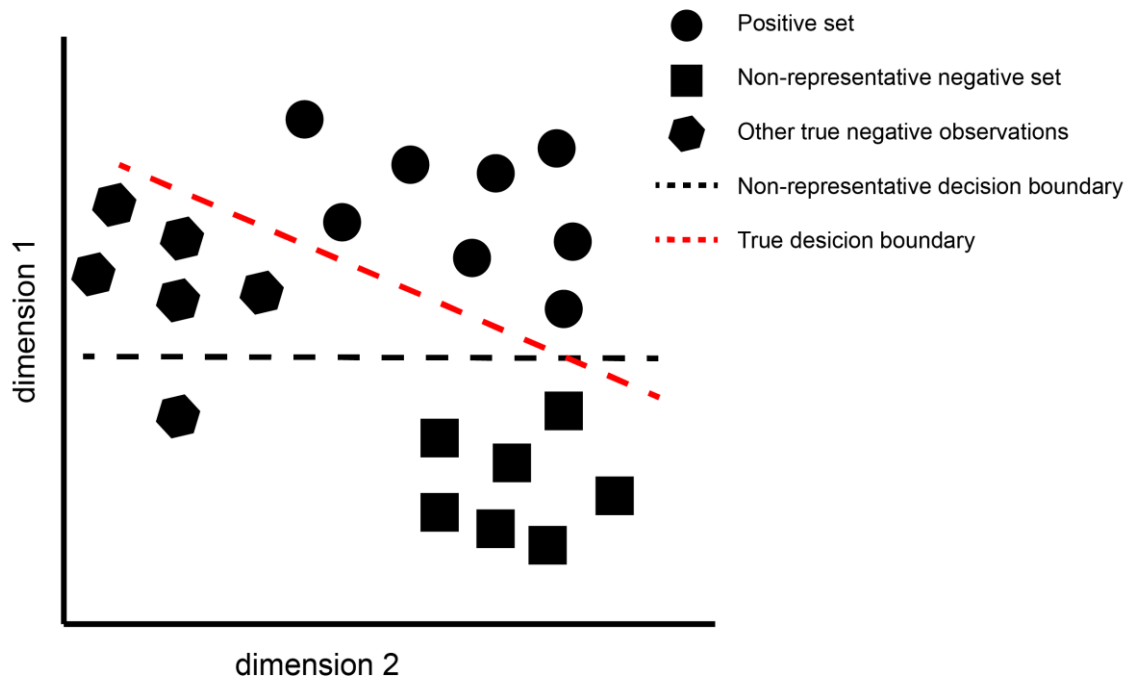


Figure 2.4. Schematic representation of two-class classification decision boundary formation in support vector machines. Non-representative set of true negatives can lead to erroneous decision boundary and inflation of false positives in the predictions.

To address the discovery of cancer genes using only positive observations in the training set, I developed a supervised classifier using a previously developed novelty detection algorithm, which was based on one-class support vector machines (Schölkopf et al. 2001). Since the method was based primarily on systems-level properties, it was named systems-level support vector machine (sysSVM). SysSVM was developed in R using the e1071 R package.

SysSVM initiated by constructing the training and prediction set for a given cohort. The training set, upon which the formation of the decision boundary was based, was comprised of all known cancer genes with damaging alterations in the samples under study. For the analysis of the pilot cohort, I considered in

total 518 known cancer genes from the Cancer Gene Census (Forbes et al. 2017). Of those, I annotated 491 (see Methods) with damaging alterations in one of the following 6 categories (see Methods):

1. Truncating mutations (stopgain, stoploss and frameshift alterations)
2. Non-truncating damaging mutations (damaging non-frameshift, nonsynonymous, splicing alterations)
3. Gain-of-function mutations
4. Homozygous deletions (Copy number = 0)
5. Heterozygous deletions with truncating or non-truncating mutations in the second allele
6. Copy number gains (Copy number ≥ 4)

The final size of the training set was comprised of 3,330 positive observations (i.e. known cancer genes harbouring damaging alterations). Consistent with the excess of genomic amplifications in oesophageal adenocarcinoma (Secrier et al. 2016; Nones et al. 2014), the vast majority of the genes (98%) in the training set were amplified (Figure 2.5A). Of the remaining 2% of the genes, only 1.5% were mutated and 0.5% were homozygously deleted. As expected, most of the known drivers were found altered in a large number of OACs. On average, each known driver was altered in 40% of samples, with a median of 7.00 and a mean of 6.78 samples (Figure 2.5B). Although the over-representation of one of the molecular features, in this case amplifications, makes the development and interpretation of the classifier challenging, the relative weight of each feature in the final model can be evaluated during the training phase. Therefore, the relative contribution of each

property can be dissected (see below). Nevertheless, the number of systems-level properties is two times higher than that of molecular features, and their contribution is expected to be significant in the final classifier, regardless of the overrepresentation of amplifications.

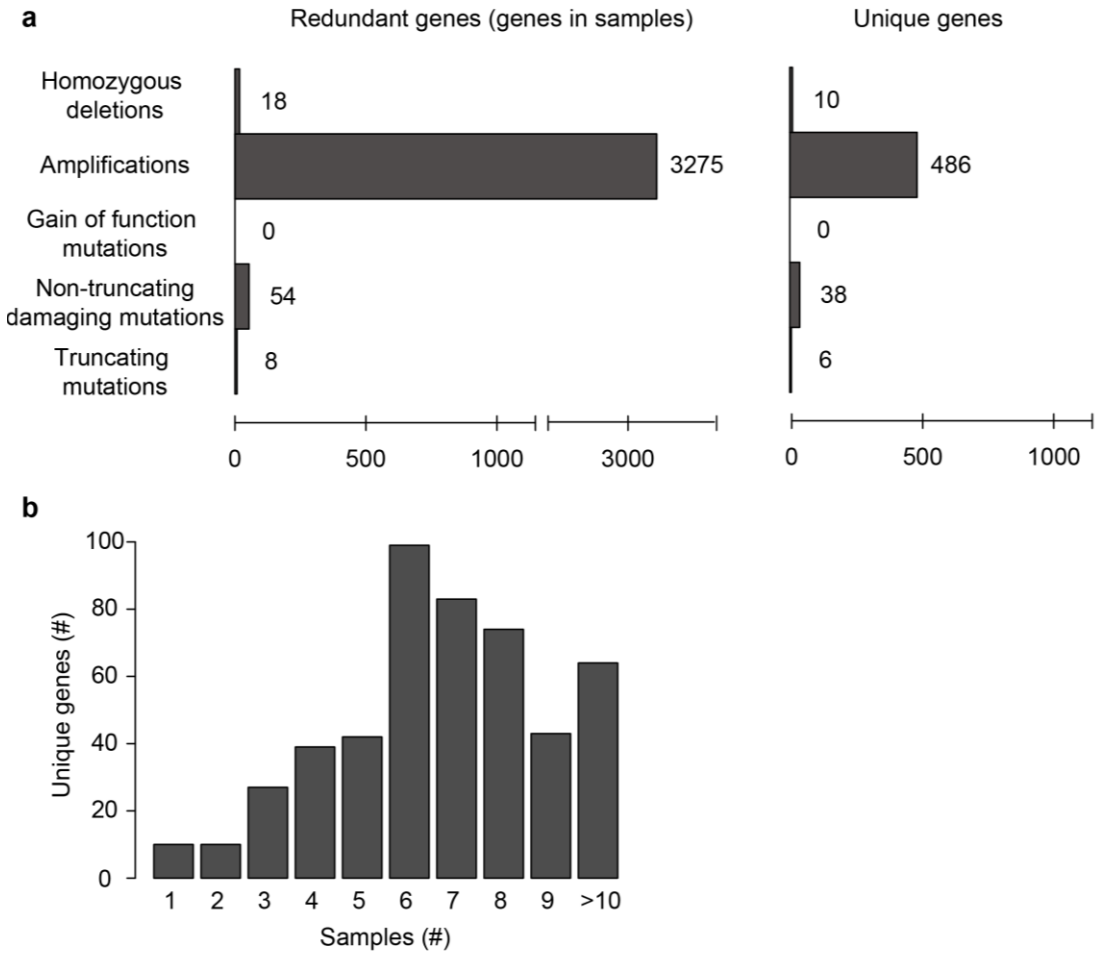


Figure 2.5. Genes in the training set of the pilot cohort. The total size of the training set was 3,330 genes (same gene in different samples counted multiple times), corresponding to 540 unique genes. **(a)** Distribution of alterations in the training set for redundant (left panel) and unique genes (right panel). **(b)** Distribution of the number of samples in which each gene was found altered.

The prediction set was comprised of any gene with damaging alterations, excluding known cancer drivers. In total, the prediction set included 18,031

unique genes, which corresponded to 121,649 redundant genes (i.e. altered genes in samples). Similar to the training set, I found that the predominant type of alteration was copy number gain and approximately 99% of genes in the prediction set belonged to this category. To account for the different numerical ranges, I scaled all continuous features (Table 2.1), in both training and prediction sets combined, to zero mean and unit variance using the following formula:

$$x_i = \frac{x_i - mean(x)}{\sigma(x)} \quad (3.9)$$

where x is the vector of values of feature *i,* which ranges from 1 to 12 for the 12 continuous features.

As already mentioned, sysSVM operates as a meta-classifier, incorporating the results of linear, radial, polynomial and sigmoid kernels. I optimised the parameters of each kernel using grid search. Depending on the kernel, I considered the following parameters:

1) *nu* (all kernels), representing the upper bound on the fraction of outliers (i.e. training genes left outside the estimated region) and the lower bound on the fraction of support vectors. Values for *nu* range from 0.05 to 0.9 with a step of 0.05 for a total of 18 values;

2) *gamma* (radial, sigmoid, and polynomial kernels), accounting for the influence of individual training points in the final model and defined as:

$$\text{gamma} = 2^x, \text{ where } x \in \{-7, -6, \ldots, 4\}$$

for a total of 18 values;

3) *degree* (polynomial kernel), representing the degree of the polynomial kernel function with three possible values (3, 4, 9);

The grid search resulted in 18 possible combinations of parameters for the linear kernel, 216 combinations for the radial and sigmoid kernels and 648 combinations for the polynomial kernel. For each combination of parameters, sysSVM performed a three-fold cross validation with 100 iterations. At each iteration, the genes of the training set were randomly split into three sub-sets, two of which (approximately 326 genes for the pilot cohort) were used for the training and the third (approximately 165 genes for the pilot cohort) as a test set. After the training was completed, the prediction was performed on the test set and the sensitivity of the model was computed for that iteration. At the end of the cross-validation, the distribution of sensitivity across all iterations was derived for each combination of parameters. The least variant (variance of sensitivity) model among the top five most sensitive models (considering the median sensitivity) in each kernel was chosen as the best model for that kernel. The resulting four best models were then trained on the whole training set and subsequently used to predict novel driver genes in the prediction set. A schematic workflow of the optimization of sysSVM, the selection of the best models, the training and prediction is shown in figure 2.6.

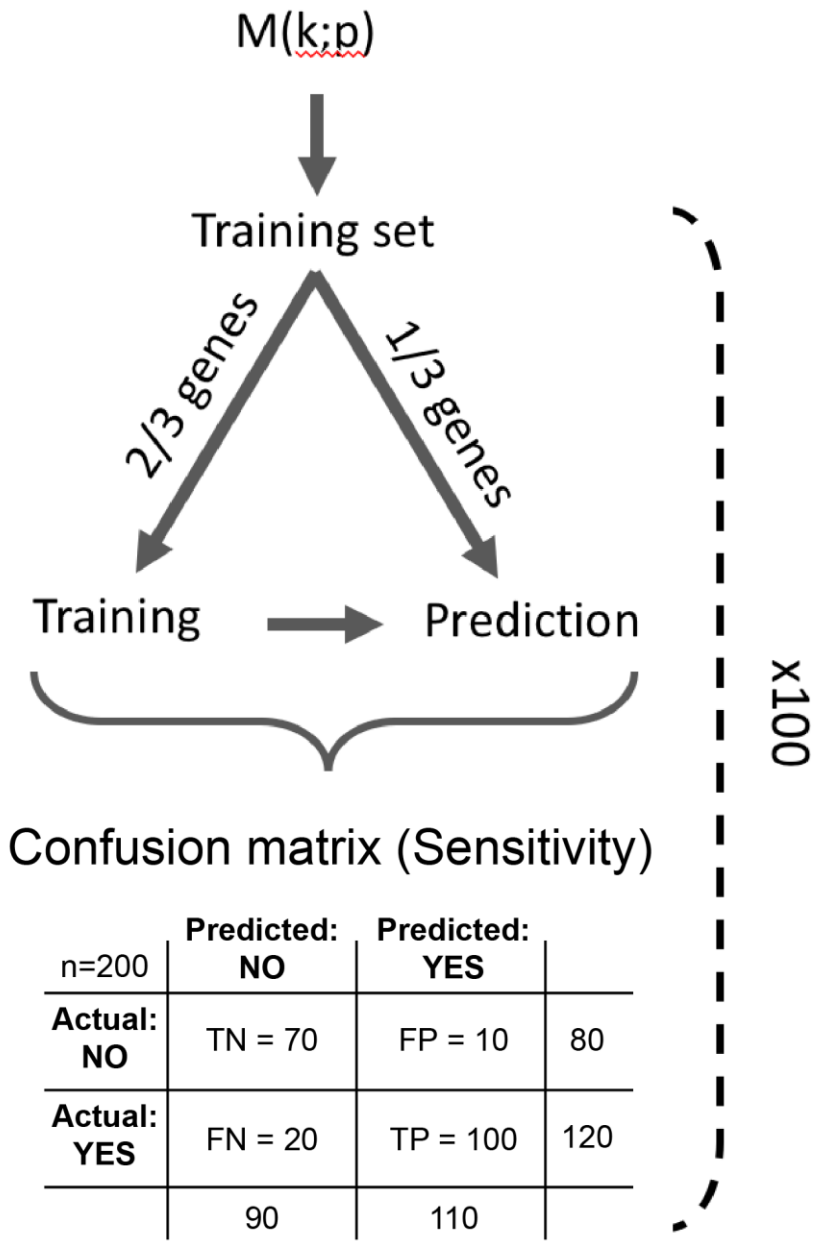


Figure 2.6. A schematic workflow of parameter optimisation using grid search in sysSVM. For a given model (M), kernel (k) and a set of parameters (p), sysSVM performs two steps. First, the genes in the training set are divided in two thirds which are used for training and one third which is used for prediction. Second, a confusion matrix is then calculated (an example is showed using 200 observations) to derive sensitivity. It is noted that sysSVM uses one-class support vector machines in which the true negative observations are absent ("Actual NO" in the schematic). Steps 1 and 2 are repeated 100 times and the best set of parameters is chosen based on the distribution and variance of the sensitivity estimates.

2.3.4 Best sysSVM models in the OAC pilot cohort

In a two-class classification setting, in which true negative observations existed, the best models would have been selected using sensitivity, specificity and other true negative-based measures (Fawcett 2006). Since I developed SysSVM based on one-class classification, all of the above-mentioned metrics, except for sensitivity, could not be computed. Therefore, I implemented a stability measure (variance of sensitivity) as an additional metric of the performance of each model during cross-validation. Apart from the polynomial kernel, models of intermediate sensitivity (ranges from 0 to 1) exhibited overall higher variance and models with the highest or lowest sensitivity were relatively stable (Figure 2.7). The selected models achieved cross-validation sensitivity higher than 85%, meaning that when novel cancer driver genes were encountered by the models, they were classified as drivers in 85% of the cases. This reflected the fact that, in the absence of true negatives, sensitivity was measured as the ratio of the number of positive predictions by sysSVM over the total size of the test set. The parameters of the best models in the pilot cohort and their corresponding sensitivity are shown in table 2.4.

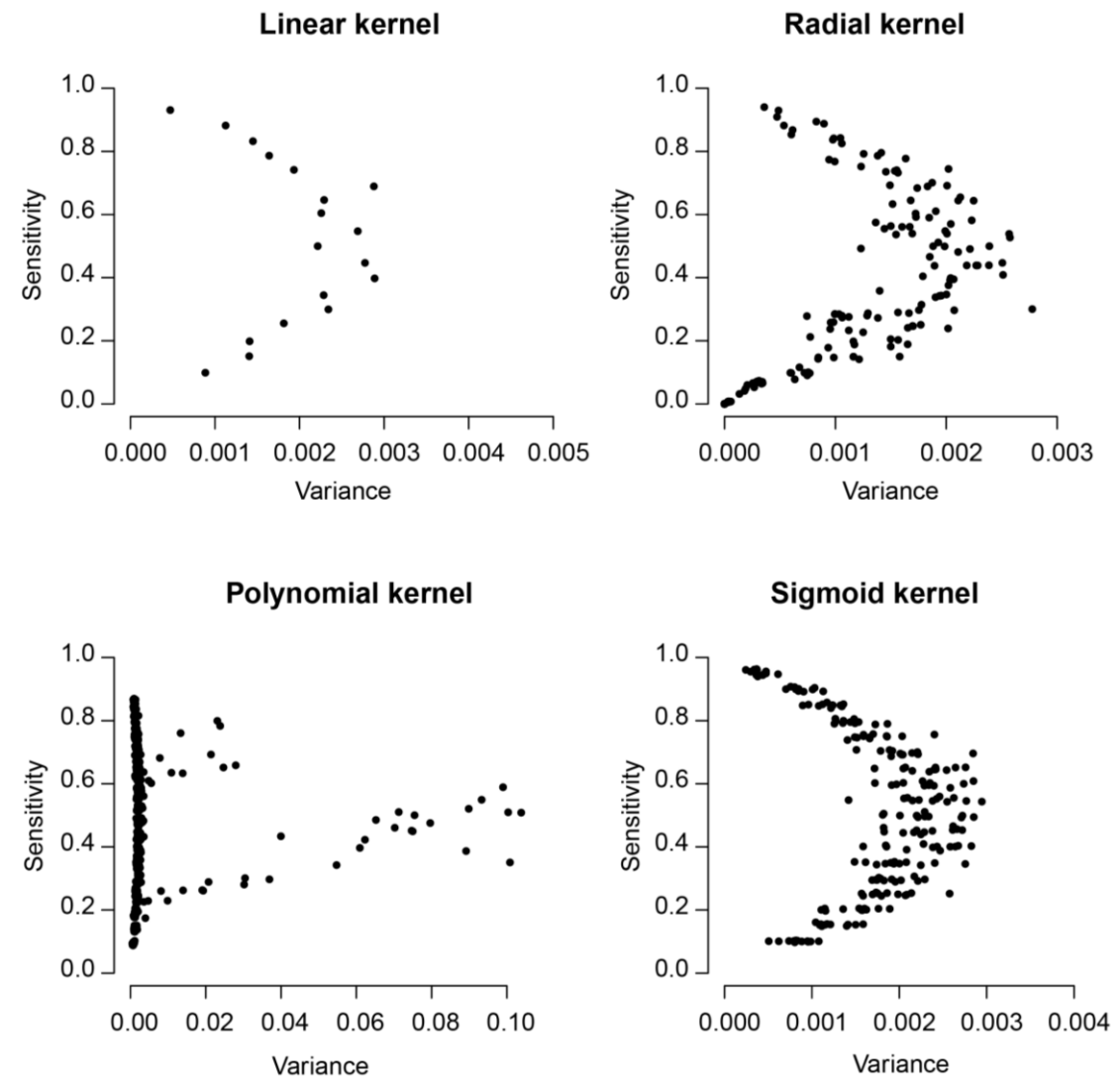


Figure 2.7. Mean sensitivity and variance for the four kernels in the pilot cohort. For each kernel, the mean sensitivity across all 100 iterations from the grid search is plotted as a function of variance. The best set of parameters was selected as the least variant among the top 5 most sensitive combinations of parameters.

All four best models had a *nu* value of 0.05, which is the minimum of the tested *nu* range. Hence, at least 95% of the training set was considered during the training phase. Higher values of *nu* would have led to the exclusion of training observations and it would indicate higher dispersion of known cancer genes in the feature space. The values of the gamma parameter were not

concentrated in a particular part of the tested range and the degree parameter was only applicable in the polynomial kernel. The number of support vectors (i.e. training observations used for the formation of the decision boundary) was comparable across the four kernels. The sigmoid kernel had the lowest number (167 genes), polynomial the highest (211 genes), and linear and radial had 181 and 172 genes, respectively. A unique feature of SVMs is that the weights of the decision function are defined by only a small subset of the training observations, the support vectors. The magnitude of the decision value can approximate how confident the model is for the classification of a certain observation. Therefore, positive values belong to positive predictions and negative values to negative predictions. Decision values can be perceived as the distance from the decision boundary. In the pilot cohort, the radial kernel had the lowest average decision value of the training observations (4.981) and the polynomial kernel had the highest (578.98), with linear and sigmoid having 54.78 and 100.35, respectively (Figure 2.8). As expected from the high sensitivity of the selected models, very few genes had negative decision values. SysSVM utilised the decision values of individual kernels to create a weighted meta-score, taking also into account the performance of each kernel (see below).

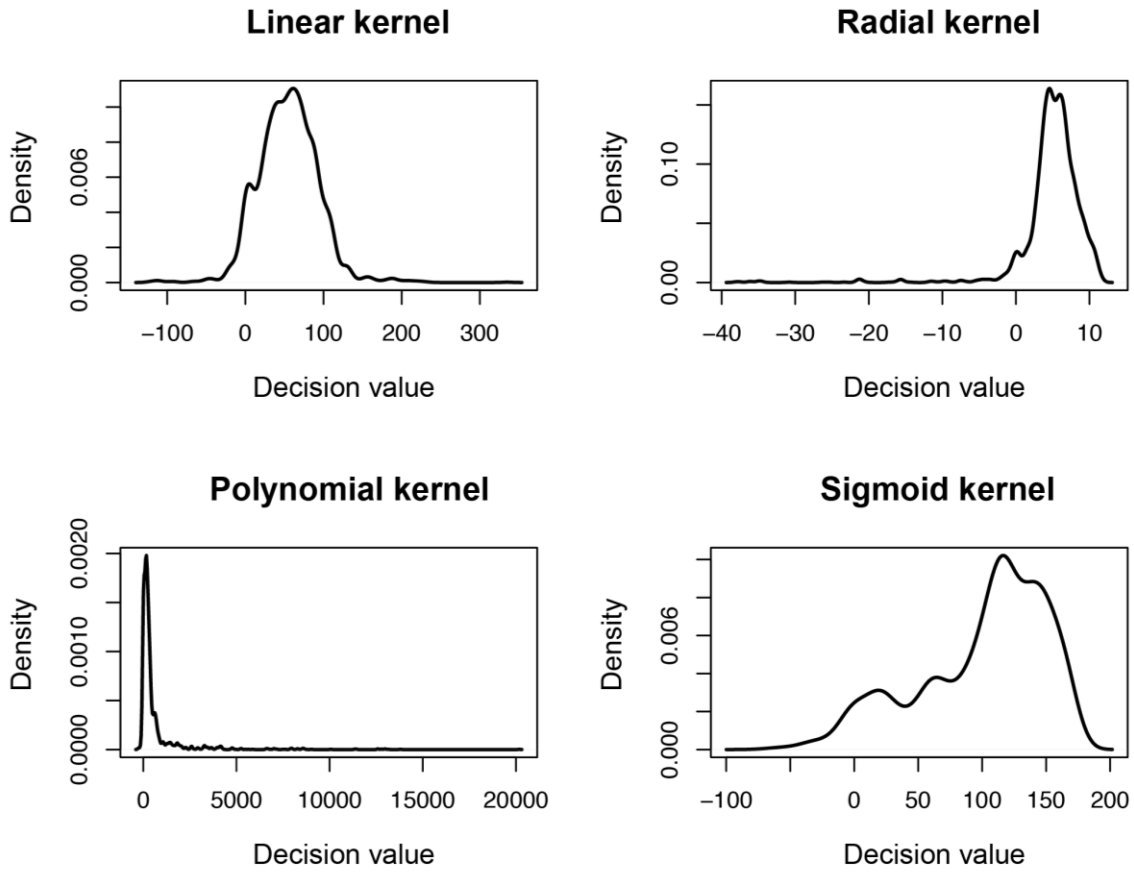


Figure 2.8. Distribution of the decision values of the best models used for prediction in the pilot cohort.

To examine which features carry the highest weight in each kernel, I employed recursive feature elimination (RFE) (Guyon et al. 2002). RFE is an iterative process in which the following three steps are repeated until all sysSVM features have been eliminated:

1. Train the classifier
2. Compute the ranking criterion
3. Remove the feature with the smallest ranking criterion and repeat

For each kernel, I used the selected parameters from cross-validation to train the classifier and computed the weight vector (w) by multiplying the sysSVM coefficients with the support vectors. Then, I defined the ranking criterion as w^2 and extracted the feature with the smallest w^2 value. Since I developed sysSVM based on 31 features, RFE run 30 iterations per kernel to estimate the ranks of the features. Copy number gains was the feature with the highest weight (rank one in three out of four kernels) and the number of gain-of-function mutations was the least important overall (Table 2.3 and Figure 2.9). This reflected the fact that I trained sysSVM using mostly amplified known cancer genes, due the high genomic instability of OAC and the number of gain-of-function mutations that were identified in the pilot cohort was zero (Figure 2.5). Overall, the molecular features were less important than the systems-level properties, except for the radial kernel in which 4 of the 7 molecular features were in the top 10 features of this classifier (Table 2.3). Besides copy number amplifications, other top-ranking features were universal expression of genes in all human tissues, whole-genome duplications, gene age and the number of connections in the protein-protein interaction network (Table 2.3). In the other end of the spectrum, the least important features were mainly molecular features, such as truncating mutations and the number of protein domains (Table 2.3 and Figure 2.9).

Table 2.3. Rank of features in individual kernels in sysSVM. Colour code represents molecular features (orange) and systems-level properties (blue).

Rank	Linear	Polynomial	Radial	Sigmoid
1	Gene gain	Gene gain	Protein degree	Gene gain
2	Medium expressed	Ubiquitously expressed	miRNA interactions	Ubiquitously expressed
3	Ubiquitously expressed	Ohnolog	Protein betweenness	Old gene
4	Ohnolog	Medium expressed	Exonic SNVs	Ohnolog
5	Old gene	Old gene	Damaging mutations	Central protein
6	Origin in vertebrates	Origin in vertebrates	Ubiquitously expressed	Hub
7	Origin in mammals	Origin in metazoans	Gene gain	Origin in eukaryotes
8	Origin in metazoans	Origin in mammals	Hub	Origin in prokaryotes
9	Selectively expressed	Selectively expressed	Central protein	Origin in metazoans
10	Gene loss	Hub	Truncating mutations	Medium expressed
11	Specifically expressed	Specifically expressed	Ohnolog	Gene duplication
12	Origin in eukaryotes	Origin in prokaryotes	Old gene	Origin in vertebrates
13	Origin in prokaryotes	Origin in eukaryotes	Origin in metazoans	miRNA interactions
14	Hub	Gene loss	Gene duplication	Origin in opisthokonts
15	Gene duplication	Gene duplication	Origin in prokaryotes	Origin in mammals
16	Central protein	Central protein	Origin in eukaryotes	Selectively expressed
17	Not expressed	Not expressed	Medium expressed	Not expressed
18	Origin in primates	Origin in primates	Origin in vertebrates	Specifically expressed
19	Origin in opisthokonts	Origin in opisthokonts	Origin in opisthokonts	Origin in primates
20	Protein degree	Chromatin state	Gene loss	Gene loss
21	miRNA interactions	Protein degree	Copy number	Protein degree
22	Gene length	miRNA interactions	Origin in mammals	Gene length
23	Number of tissues	Gene length	Selectively expressed	Number of tissues
24	Damaging mutations	Number of tissues	Not expressed	Protein betweenness
25	Chromatin state	Protein betweenness	Specifically expressed	Damaging mutations
26	Exonic SNVs	Exonic SNVs	Origin in primates	Exonic SNVs
27	Copy number	Damaging mutations	Gene length	Truncating mutations
28	Protein betweenness	Protein domains	Protein domains	Copy number
29	Protein domains	Copy number	Number of tissues	Chromatin state
30	Gain-of-function mutations	Truncating mutations	Chromatin state	Protein domains
31	Truncating mutations	Gain-of-function mutations	Gain-of-function mutations	Gain-of-function mutations

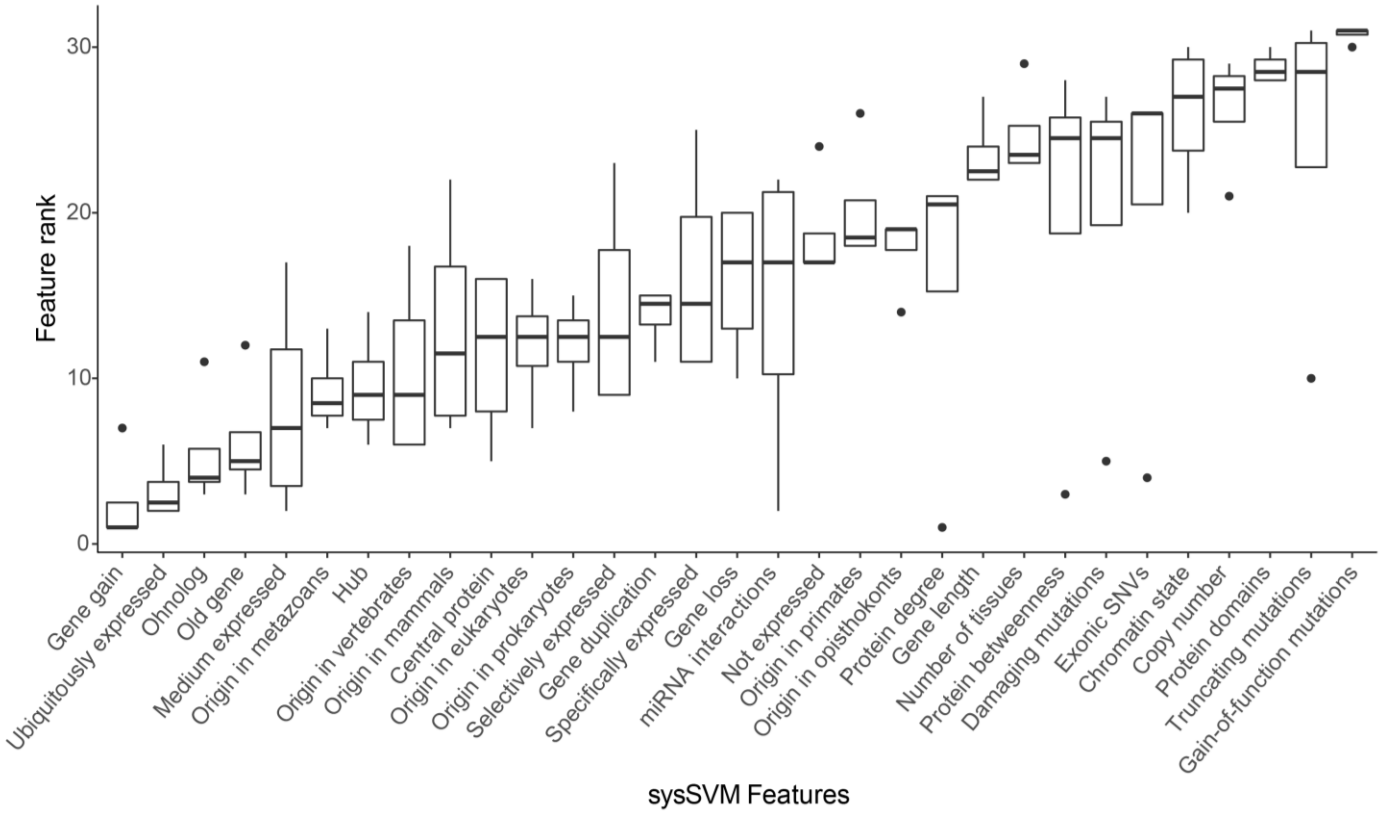


Figure 2.9. Summary of ranks retrieved by recursive feature elimination across all kernels implemented in sysSVM. Boxplots are sorted based on median rank.

Table 2.4. Parameters and performance of the four best models used to predict cancer genes in the pilot cohort. After grid search with all possible combination of parameters (18 for the linear kernel, 216 for the radial and sigmoid kernels and 648 for the polynomial kernel), the best model in each of the four kernels were chosen based on the median and variance of the sensitivity distribution across 100 iterations of cross-validation.

Kernel	nu	Gamma	Degree	Cross-validation sensitivity			
				Minimum	Median	Maximum	Variance
Linear	0.05	NA	NA	0.87	0.94	0.98	0.0005
Polynomial	0.05	0.5	3	0.78	0.87	0.93	0.0008
Radial	0.05	0.0078	NA	0.89	0.94	0.98	0.0004
Sigmoid	0.05	8	NA	0.92	0.96	1	0.0002

2.3.5 Formulation of sysSVM meta-score

After I optimised the parameters, I trained the selected best model in each kernel on the whole training set. Subsequently, sysSVM constructed a meta-classifier by combining the decision values of each kernel into a single score. I hypothesised, that although some kernels performed better than others (as measured by the sensitivity), it was not warranted that all regions of the features space were described equally well by the single best kernel. Therefore, a multi-kernel approach could predict genes that would not have been visible using a single kernel, although each kernel should have contributed proportionally to its sensitivity to the final score. To this end, the formulation of the score equation was designed as a weighted average of individual kernels in which the sensitivity of each kernel was used as a weighting factor. Hence, sub-optimal kernels contributed to the final score of each gene but disproportionally to the more sensitive kernels. SysSVM computed a combined score (S_{gs}) for each altered gene *(g)* in each sample *(s)* as follows:

$$S_{gs} = \frac{\sum_{i=1}^{4} \left(-log_{10}\left(\frac{R_{igs}}{N_s}\right) \times BMS_i\right)}{4 \times log_{10}(N_s)} \qquad (2.10)$$

where N is the number of altered genes in sample s; R_{igs} is the rank of gene g in sample s and kernel i; and BMS_i is the sensitivity of the best model in kernel i. R_{igs} is derived by sorting the decision values (indicative of the distance of the gene from the decision boundary) of kernel i within sample s so that high decision values correspond to top scoring genes.

The combined score S_{gs} corrected for the total number of altered genes in that sample (N_s) and for the sensitivity of each kernel and applied a

normalisation factor (denominator of equation 2.10) to scale the resulting value between 0 and 1. Overall, genes that were predicted by higher number of kernels had higher scores (Figure 2.10). This suggested that the combined score sufficiently captured both the majority rule and the distance from the boundary of each kernel.

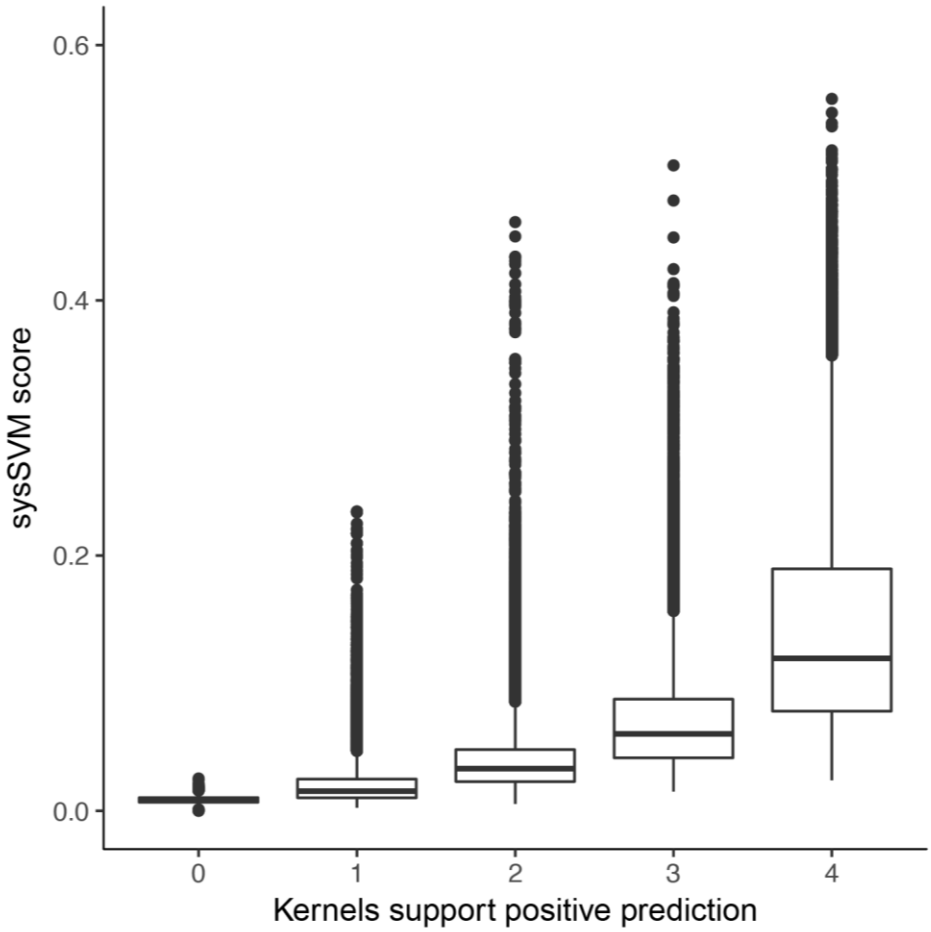


Figure 2.10. Distribution of sysSVM score for 121,649 genes in the prediction set of the pilot cohort as a function of kernels supported the positive prediction of each gene.

2.3.6 Relevance of sysSVM predictions to OAC pathogenesis

In the initial stages of the development of sysSVM, it was essential to manually inspect genes and pathways in order to analyse the predictions of cancer genes in the pilot cohort and their relevance to OAC pathogenesis. Moreover, pathway information and limited experimental validation could feedback to the relevance of predictive features used in sysSVM and their potential to identify new driver genes in OAC. To this end, I selected the top 10 scoring genes in each sample (88 genes) and performed pathway enrichment analysis using the Reactome database (Milacic et al. 2012). As expected, the majority of the predicted genes (70%) were altered in only one sample, whereas only 13% of them was recurrently altered in 5 samples or more (Table 2.5). From the 1,156 pathways that I tested (see Methods), 53 were found enriched (FDR<0.01; Figure 2.11). Given the limited number of samples in the pilot cohort, the primary focus of this analysis was the inspection and validation of some of the predictions, rather than the identification of novel pathways in OAC. For the latter, I applied sysSVM to a larger cohort of OACs (see chapter 3).

I found that numerous genes that were predicted by sysSVM contributed to the enrichment of several pathways, whose implication to tumorigenesis and DNA damage response has been previously described. In particular, transcriptional regulation by *TP53*, Notch-related pathways, *MAPK* signalling, regulation of DNA replication and DNA damage checkpoints were all found enriched with several genes and samples involved in each one of them (Figure 2.11). A particularly interesting test case was that of *TP53*-related pathways, since predicted genes in these pathways could have been potential true positives. In OAC, around 70% of patients have been reported to harbour mutations in *TP53* and this was consistent in our pilot cohort, in which 13/18

(72%) samples had predicted damaging mutation in *TP53*. SysSVM scored high 9 genes involved in *TP53*-related pathways in 14 samples. Of interest, sysSVM predicted interactors of *TP53* in samples with wild-type *TP53*, suggesting that more samples had tumorigenic alterations in *TP53*-related pathways than those that were identified using only the *TP53* status. In particular, four samples with wild-type *TP53* had amongst their top 10 scoring genes *AGO1* or *AGO2*, both of which are key components of the RNA-induced silencing complexes (RISCs) and have been previously linked to DNA damage response and miRNA-mediated gene transcriptional modulation (Krell et al. 2016).

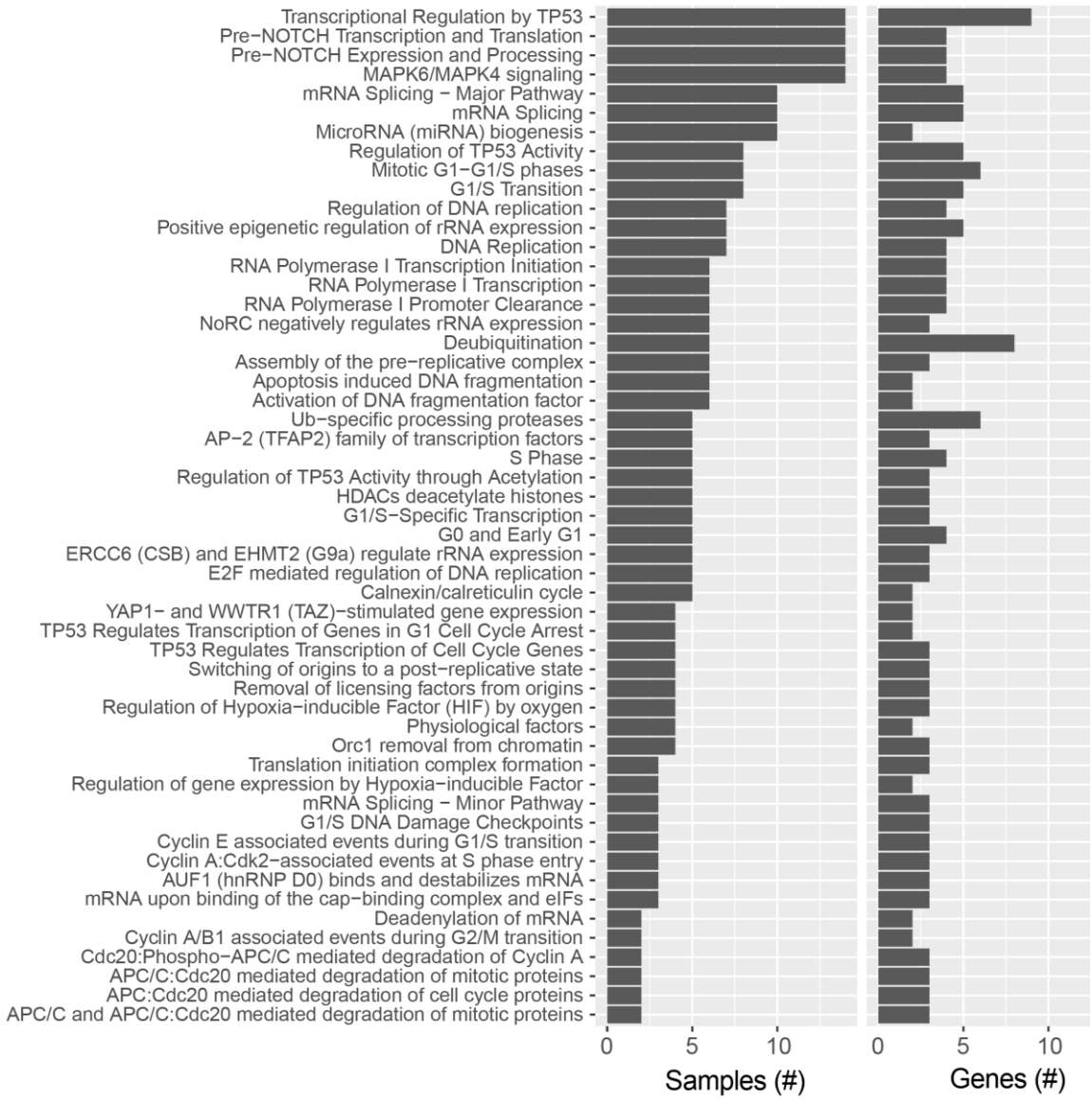


Figure 2.11. Pathway enrichment analysis of 88 sysSVM top 10 scoring genes in the pilot cohort. For each one of the 53 pathways found enriched, the number of samples and genes contributed to the enrichment are reported.

Another strong enrichment of sysSVM predictions was found in pathways related to Notch signalling. Given that Notch acts as an oncogene and its contribution to tumorigenesis of OAC has been previously reported (Wang et al. 2014), I examined the genes involved in Notch-related pathways. Apart from *AGO1* and *AGO2*, which belong to both *TP53* pathways and Notch signalling

according to REACTOME, *KAT2B*, a histone acetyl-transferase, and *E2F1*, a transcription factor, whose expression is increased in many cancer types, were also responsible for the enrichment.

The human *KAT2B*, and its 70% identical paralogue *KAT2A*, catalyse lysine acetylation, which is the process of transferring the acetyl group of acetyl-CoA to the epsilon amino group of lysing residues (Yang and Seto 2008). As histone acetyl-transferases (HATs), *KAT2A* and *KAT2B* are members of a family of ~20 enzymes (Sadoul et al. 2011), which are involved in lysine acetylation and post-transcriptional modification of more than 6,000 proteins (Hornbeck et al. 2012). Through the acetylation of histones and other proteins, KATs play an essential role in the modulation of transcription activation and DNA replication and repair (Espinosa et al. 2010; Kelly et al. 2009; Pai et al. 2014; Orpinell et al. 2010). This makes them particularly interesting targets for further experimental validation of their driver role in OAC. Apart from histones, *KAT2A* and *KAT2B* can also acetylate non-histone targets, such as *CDC6* and cyclin A, both of which regulate G1/S cell cycle transition and mitosis (Paolinelli et al. 2009), a process that is dysregulated via multiple sysSVM predictions in a significant fraction of OACs (see chapter 4). Using expression data from the Xena browser (https://xenabrowser.net/) for a cohort of more than 11,000 cancer samples from TCGA, both *KAT2B* and *KAT2A* expression were found associated with *E2F1* expression. However, *KAT2B* showed a negative association (Figure 2.12A), while *KAT2A* expression was positively associated with *E2F1* expression levels (Figure 2.12B), suggesting a differential functional dependency of the two genes to *E2F1*. Unlike *KAT2B*, the functional link of *E2F1* and *KAT2A* has been reported before (Chen et al. 2013). It is intriguing that a proapoptotic transcription factor, such as *E2F1*, regulates the expression

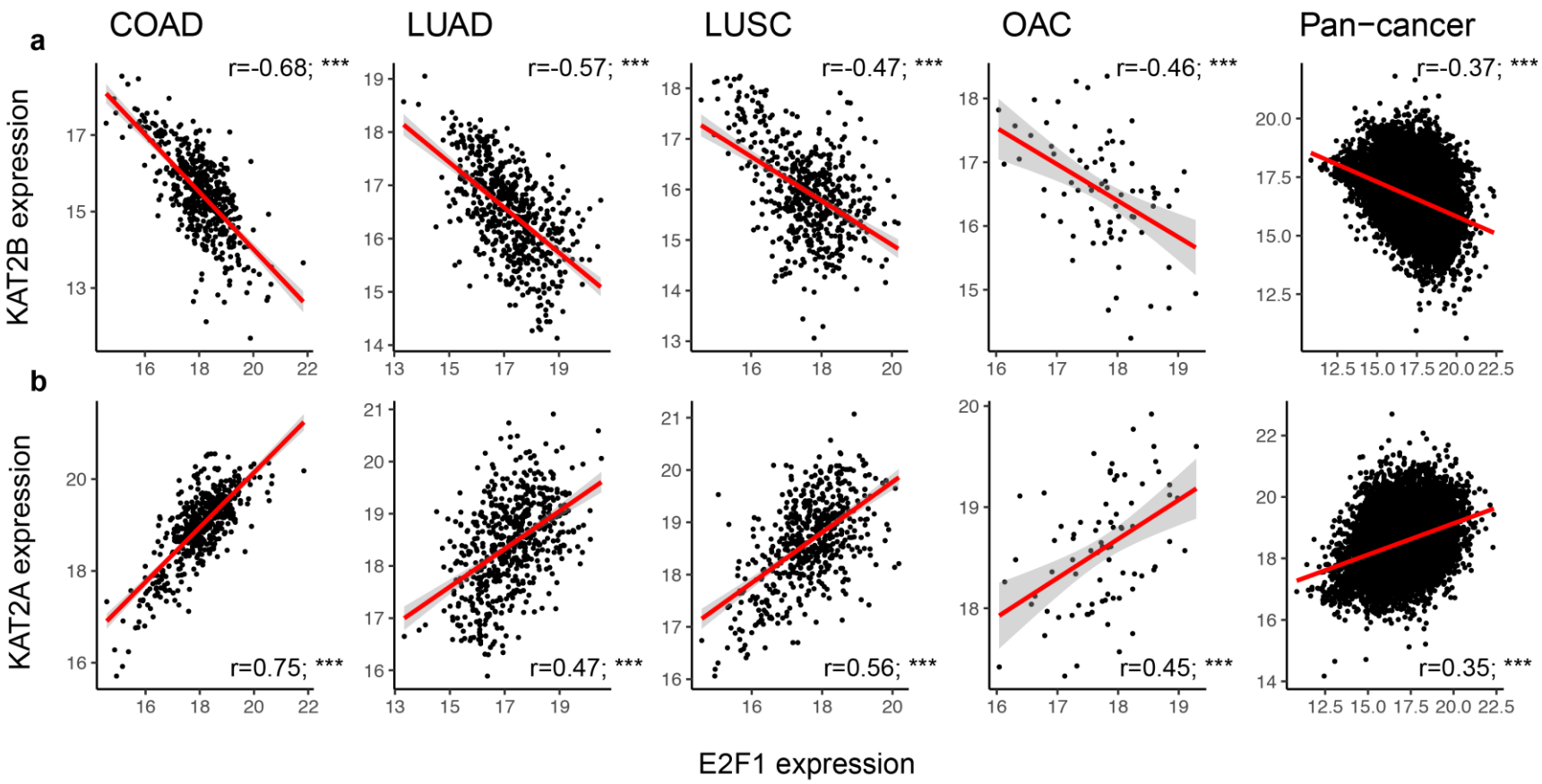
of an antiapoptotic gene, such as *KAT2A*. The seemingly opposite functions of *KAT2A* and *E2F1* highlight a negative functional feedback in which *E2F1*-induced cell death is antagonised by overexpression of *KAT2A* to promote cancer progression (Farria, Li, and Dent 2015). Apart from colon and lung cancer, where this dependency has been demonstrated before, our data suggest that *E2F1* and KATs might be functionally connected also in OAC.

To examine the prevalence of *KAT2A* and *KAT2B* somatic alterations in multiple cancer types, I performed a pan-cancer analysis of 7,828 cancer samples from 31 cancer types (Table 1.1), which revealed 1,196 samples (15.3% of the total) with non-silent mutations and deletions. Overall, 25 cancer types had damaging modifications in >5% of the samples (Figure 2.12C). In particular, I found that more than 35% of sequenced samples in kidney chromophobe, ovarian serous cystadenocarcinoma, uveal melanoma, cholangiocarcinoma and kidney renal clear cell carcinoma harboured modifications in these genes. The most frequent type of alteration was gene loss (91%), followed by nonsynonymous SNVs (6.5%) (Figure 2.12D). In the OAC pilot cohort, *KAT2B* harboured predicted damaging alterations in six OACs, while in one sample was ranked in the top 10 scoring genes. In the remaining five OACs, *KAT2B* was ranked as the 11th, 12th, 15th, 21st and 33rd, very close to the top 10 in at least three of the five. Additionally, *E2F1* was predicted by sysSVM in three OACs (Table 2.5).

To test whether alterations of KATs, and in particular of *KAT2A* and *KAT2B*, leads to a cancer-related phenotype, we[1] knocked-out both genes in OE19 oesophageal cancer cell line using a vector-free CRISPR system that was previously developed in our lab (Benedetti et al. 2017). Briefly, three

[1] Experiments were performed by Dr. Lorena Benedetti

pooled crRNAs were co-transfected with Cas9 and tracrRNA in OE19 cells (Table 2.6). We then verified that edited OE19 cells exhibited increased proliferation (Figure 2.12E) and decreased apoptosis as compared to wild-type cells (Figure 2.12F). These data suggest that *KAT2A* and *KAT2B* upon loss of function modification contribute to tumorigenesis in OAC and, therefore, were true positive predictions of sysSVM.


a
b
COAD
LUAD
LUSC
OAC
Pan−cancer
KAT2B expression
KAT2A expression
E2F1 expression
r=-0.68; ***
r=-0.57; ***
r=-0.47; ***
r=-0.46; ***
r=-0.37; ***
r=0.75; ***
r=0.47; ***
r=0.56; ***
r=0.45; ***
r=0.35; ***

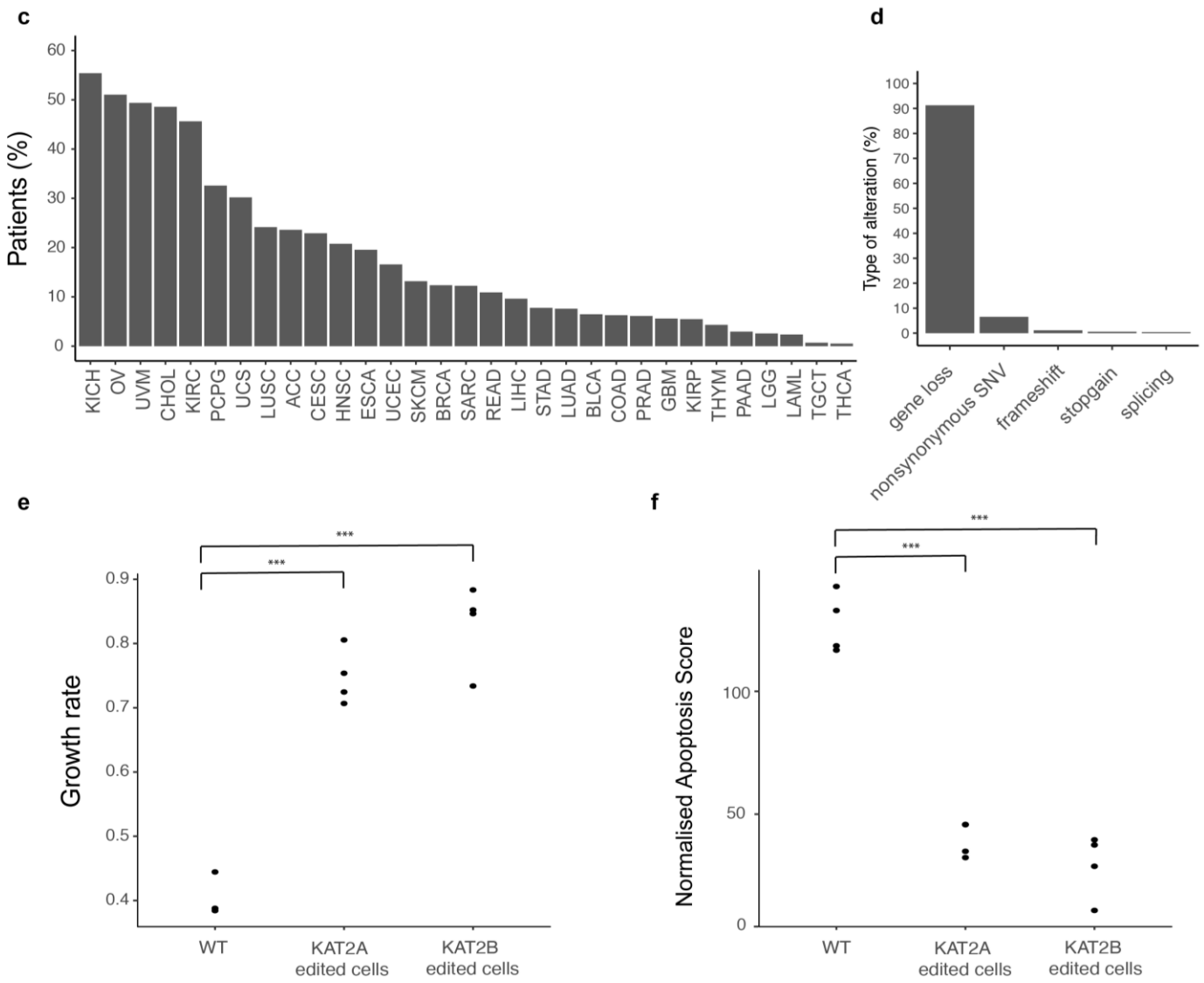

c
Patients (%)
0
10
20
30
40
50
60
KICH
OV
UVM
CHOL
KIRC
PCPG
UCS
LUSC
ACC
CESC
HNSC
ESCA
UCEC
SKCM
BRCA
SARC
READ
LIHC
STAD
LUAD
BLCA
COAD
PRAD
GBM
KIRP
THYM
PAAD
LGG
LAML
TGCT
THCA
d
Type of alteration (%)
0
10
20
30
40
50
60
70
80
90
100
gene loss
nonsynonymous SNV
frameshift
stopgain
splicing
e
Growth rate
0.4
0.5
0.6
0.7
0.8
0.9
***
***
WT
KAT2A edited cells
KAT2B edited cells
f
Normalised Apoptosis Score
0
50
100
***
***
WT
KAT2A edited cells
KAT2B edited cells

Figure 2.12. Pan-cancer analysis and experimental validation of *KAT2A* and *KAT2B* acetyl-transferases. Correlation of expression levels of **(a)** *KAT2A* and **(b)** *KAT2B* with *E2F1* expression in colon adenocarcinoma (COAD), lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), oesophageal adenocarcinoma (OAC) and pan-cancer cohorts from TCGA. **(c)** Percentage of patients with non-silent mutations or gene deletions (i.e. heterozygous or homozygous copy number aberrations) in 7,828 samples from TCGA. Abbreviations and number of samples per cancer type are reported in table 1.1. **(d)** Most prevalent types of alterations of *KAT2A* and *KAT2B* in the pan-cancer cohort. Gene knock-out of *KAT2A* and *KAT2B* was performed on OE19 oesophageal cancer cell line, using a vector-free CRISPR-mediated editing approach as previously described (Benedetti et al. 2017) (see Methods)**. (e)** For proliferation, four replicates per condition were measured every three hours up to 75 hours in Incucyte Live Cell analysis system. Growth rate was calculated for each condition by fitting initially a sigmoid model in the data and subsequently a linear model in exponential phase of each experiment. An average growth rate was calculated across the four replicates. Conditions were compared using the two-tailed Student's t-test. **(f)** For the same conditions as in e, apoptosis was estimated by measuring the caspase activity in Incucyte Live Cell analysis system. The normalised apoptosis score was derived as the ratio of the caspase activity measurement in each experiment over its confluence, to account for slight deviations of the proliferation rate across experiments. Three asterisks denote p value < 0.001. The full names of cancer types are reported in table 1.1.

Table 2.5. SysSVM top 10 scoring predictions in each OAC in the pilot cohort. For each gene the number of OACs in which it was predicted as driver, the average sysSVM score and the kernels predicted its driver role are reported.

Gene symbol	OACs (#)	Average sysSVM score	Kernels with positive prediction
NCOA3	11	0.471348555	linear, radial, polynomial, sigmoid
TRIM28	11	0.476008422	linear, radial, polynomial, sigmoid
BAG6	10	0.446418377	linear, radial, polynomial, sigmoid
HNRNPUL1	8	0.444488626	linear, radial, polynomial, sigmoid
AGO2	6	0.434619329	linear, radial, polynomial, sigmoid
DNMT1	5	0.441461596	linear, radial, polynomial, sigmoid
GANAB	5	0.42215355	linear, radial, polynomial, sigmoid
KPNB1	5	0.436417317	linear, radial, polynomial, sigmoid
LARP1	5	0.437587079	linear, radial, polynomial, sigmoid
POGZ	5	0.432702271	linear, radial, polynomial, sigmoid
SPTBN1	5	0.425580056	linear, radial, polynomial, sigmoid
AGO1	4	0.428647663	linear, polynomial
ATN1	4	0.433762497	linear, radial, polynomial, sigmoid
ASAP1	3	0.458644783	linear, radial, polynomial, sigmoid
COL4A1	3	0.428521467	linear, radial, polynomial, sigmoid

E2F1	3	0.461876384	linear, radial, polynomial, sigmoid
HIPK2	3	0.436336561	linear, radial, polynomial, sigmoid
NOP2	3	0.432275619	linear, radial, polynomial, sigmoid
TXLNA	3	0.460353624	linear, radial, polynomial, sigmoid
CHD4	2	0.454131694	linear, radial, polynomial, sigmoid
ERBB2IP	2	0.426749082	linear, radial, polynomial, sigmoid
FLT1	2	0.491169343	linear, radial, polynomial, sigmoid
GATAD2B	2	0.441963149	linear, radial, polynomial, sigmoid
GLI1	2	0.439110303	linear, radial, polynomial, sigmoid
HIF1A	2	0.438197968	linear, radial, polynomial, sigmoid
MCM7	2	0.466751676	linear, radial, polynomial, sigmoid
PRRC2A	2	0.432629967	linear, radial, polynomial, sigmoid
SIN3B	2	0.441902237	linear, radial, polynomial, sigmoid
ACACA	1	0.508598838	linear, radial, polynomial, sigmoid
ACTN4	1	0.505733986	linear, radial, polynomial
AKAP11	1	0.489678272	linear, radial, polynomial, sigmoid
APPL1	1	0.503229099	linear, radial, polynomial, sigmoid
ATP7B	1	0.546960402	linear, radial, polynomial, sigmoid
ATP8B2	1	0.451790665	linear, radial, polynomial, sigmoid
CALU	1	0.413428841	linear, radial, polynomial, sigmoid
CANX	1	0.410961215	linear, radial, polynomial, sigmoid
CCNA1	1	0.457717273	linear, radial, polynomial, sigmoid
CDC20	1	0.428716654	linear, radial, polynomial, sigmoid
CDC25A	1	0.498309324	linear, radial, polynomial, sigmoid
CDK19	1	0.486220951	linear, radial, polynomial, sigmoid
CTBP2	1	0.424220966	linear, radial, polynomial, sigmoid
CUX1	1	0.423540355	linear, radial, polynomial, sigmoid
DDX17	1	0.44619495	linear, radial, polynomial, sigmoid
DHX30	1	0.441318096	linear, radial, polynomial, sigmoid
DNMT3B	1	0.421640013	linear, radial, polynomial, sigmoid
EEF2	1	0.424448845	linear, radial, polynomial
EIF4G1	1	0.423686843	linear, radial, polynomial, sigmoid
FYCO1	1	0.461600916	linear, radial, polynomial, sigmoid
HMGB1	1	0.51535323	linear, radial, polynomial, sigmoid
HNRNPL	1	0.514180501	linear, radial, polynomial, sigmoid
HSPH1	1	0.557846097	linear, radial, polynomial, sigmoid
KAT2B	1	0.42817236	linear, radial, polynomial, sigmoid
KATNAL1	1	0.502438335	linear, radial, polynomial, sigmoid
KDM5B	1	0.40509567	linear, radial, polynomial, sigmoid
KIF2C	1	0.426583094	linear, radial, polynomial, sigmoid
LATS2	1	0.484914521	linear, radial, polynomial, sigmoid
MAP4	1	0.455923164	linear, radial, polynomial, sigmoid
MMP9	1	0.44220687	linear, radial, polynomial, sigmoid
MTA2	1	0.411317798	linear, radial, polynomial, sigmoid

MYO1C	1	0.479637608	linear, radial, polynomial, sigmoid
OXSR1	1	0.442391667	linear, radial, polynomial, sigmoid
PABPC1	1	0.412361002	linear, radial, polynomial, sigmoid
PEX5	1	0.432331984	linear, radial, polynomial, sigmoid
PITPNA	1	0.425100624	linear, radial, polynomial, sigmoid
PRPF8	1	0.511162116	linear, radial, polynomial, sigmoid
PSMD11	1	0.421977785	linear, radial, polynomial, sigmoid
PUM1	1	0.478116616	linear, radial, polynomial, sigmoid
RASA1	1	0.461295249	linear, polynomial
RBL1	1	0.420673799	linear, radial, polynomial, sigmoid
RBM10	1	0.440777578	linear, radial, polynomial, sigmoid
RIMS2	1	0.450053941	linear, polynomial
RPS2	1	0.478204533	linear, radial, polynomial
RTN4	1	0.40329935	linear, radial, polynomial, sigmoid
SF3B3	1	0.427319113	linear, radial, polynomial, sigmoid
SKI	1	0.410332875	linear, radial, polynomial, sigmoid
SLC7A1	1	0.478120625	linear, radial, polynomial, sigmoid
SMARCC1	1	0.461495874	linear, radial, polynomial, sigmoid
STX12	1	0.447389895	linear, radial, polynomial, sigmoid
TGFBR2	1	0.473708203	linear, radial, polynomial, sigmoid
TRAPPC3	1	0.439131231	linear, radial, polynomial, sigmoid
UBA1	1	0.444941414	linear, radial, polynomial, sigmoid
USP19	1	0.455257781	linear, radial, polynomial, sigmoid
VAPB	1	0.437985891	linear, radial, polynomial, sigmoid
VEGFA	1	0.433731516	linear, polynomial
WASF1	1	0.453245359	linear, radial, polynomial, sigmoid
WNK1	1	0.437629951	linear, radial, polynomial, sigmoid
XPO7	1	0.413108333	linear, radial, polynomial, sigmoid
YBX1	1	0.449267139	linear, radial, polynomial

Table 2.6. Oligos used to knock-out *KAT2A* and *KAT2B* via CRISPR

Gene	Oligo	Sequence
KAT2A	KAT2A_crRNA1	GCTTTCGGCCAATGCGGCCCGG
	KAT2A_crRNA2	TATACTCCTTAGGCATGCGCGG
	KAT2A_crRNA3	CCAAGCGGCTCCGTGTGATGGG
KAT2B	KAT2B_crRNA1	GTTCTGCGACAGTCTACCTCGG
	KAT2B_crRNA2	GTTATGAGGCGACAACTCCTGG
	KAT2B_crRNA3	ATCGCAGTCTTCGTTGAGATGG

2.3.7 Comparison of sysSVM predictions to those of other methods

To further evaluate the predictions of sysSVM, I compared the top 10 scoring genes in the pilot cohort to the cancer genes predicted by other methods, such as IntOGen (Gonzalez-Perez et al. 2013) and Hotnet2 (Leiserson et al. 2015). These are the only methods that were designed to predict rare cancer drivers and, therefore, they were selected as the most appropriate ones to be compared with sysSVM. IntOGen was run *vi*a its web interface (https://www.intogen.org/analysis/home), and Hotnet2 as described in the corresponding manual (https://github.com/raphael-group/hotnet2). The default parameters were used for both tools. The input used for both methods was that of sysSVM, i.e. genes with predicted damaging alterations, as described above. IntOGen predicted 19 significant genes (Table 2.7), six of which were known cancer genes and two had been previously reported to play a role in tumorigenesis (An et al. 2016). Hotnet2 predicted 82 genes (Table 2.8), seven of which were known cancer genes and 24 had been previously linked to tumorigenesis (An et al. 2016). As sysSVM used known cancer genes for training, and therefore, no prediction was performed, those were excluded from the comparison. No overlap was observed with the 13 IntOGen genes and only 2 genes (*ERBB2IP* and *RIMS2*) were also predicted by HotNet2, while the overlap between IntOGen and HotNet2 was 3 genes (*ERBB4*, *CACNA1A* and *PTPN13*). The low overlap between the predictions of all three tools suggests different and possibly complementary principles for the discovery of cancer genes. It also highlights the difficulties in identifying rare cancer genes when deviating from recurrence-based approaches.

Table 2.7. IntOGen predictions (n=19) in the pilot cohort. For each gene that was predicted as cancer gene by IntOGen in the pilot cohort, the ensembl gene id, the gene symbol, the cancer involvement (Yes = Known cancer gene; No = Not involved in tumorigenesis; Maybe = limited evidence), the functional bias p value before and after correction, the sample frequency and the total number of samples used for the analysis are reported.

Gene	Symbol	Cancer involvement	FM p value	FM q value	Number of samples	Proportion of samples
ENSG00000072121	ZFYVE26	No	0.01178255	0.190663083	3	0.166666667
ENSG00000168036	CTNNB1	Yes	0.01942333	0.230490187	10	0.555555556
ENSG00000163629	PTPN13	No	0.049575556	0.460289999	14	0.777777778
ENSG00000172296	SPTLC3	No	0.016026343	0.218932085	9	0.5
ENSG00000141646	SMAD4	Yes	5.47076E-05	0.003245987	4	0.222222222
ENSG00000178568	ERBB4	Maybe	0.017219378	0.218932085	18	1
ENSG00000115414	FN1	No	0.007474242	0.179945302	12	0.666666667
ENSG00000141837	CACNA1A	No	0.009072385	0.179945302	17	0.944444444
ENSG00000134982	APC	Yes	0.000894472	0.039803989	9	0.5
ENSG00000137124	ALDH1B1	No	0.022633011	0.251792243	2	0.111111111
ENSG00000047410	TPR	Yes	0.014909216	0.218932085	5	0.277777778
ENSG00000160145	KALRN	Maybe	0.034316715	0.359316187	16	0.888888889
ENSG00000020633	RUNX3	No	0.009599619	0.179945302	5	0.277777778
ENSG00000062370	ZFP112	No	0.008659662	0.179945302	6	0.333333333
ENSG00000159307	SCUBE1	No	0.010109287	0.179945302	7	0.388888889
ENSG00000073331	ALPK1	No	0.008687294	0.179945302	6	0.333333333
ENSG00000176102	CSTF3	No	0.04259737	0.421240658	9	0.5
ENSG00000141510	TP53	Yes	0	0	15	0.833333333
ENSG00000147889	CDKN2A	Yes	1.31095E-05	0.001166741	4	0.222222222

Table 2.8. HotNet2 predictions (n=82) in the pilot cohort. For each gene that was predicted as cancer gene by HotNet2 in the pilot cohort, the gene symbol and the cancer involvement (Yes = Known cancer gene; No = Not involved in tumorigenesis; Maybe = limited evidence) are reported.

Gene symbol	Cancer involvement
CDH11	Yes
CDH2	No
CDH4	No
CNKSR2	No
CTNND2	No
DLG2	No
DSCAML1	No
ERBB2	Yes
ERBB2IP	Maybe
ERBB3	Maybe
ERBB4	Maybe
ERRFI1	Maybe
FZD4	No
GRIK2	No
GRIK4	No
KCNMA1	No
MAGI2	No
MAP1A	No
MAP1B	No
NRG3	Maybe
SCN5A	Maybe
SNTB1	No

Gene symbol	Cancer involvement
AKAP6	Maybe
CACNA1A	No
CACNA1C	No
CACNA1S	No
CACNB4	No
ERC2	No
PCLO	Maybe
PDE4DIP	Yes
PPM1A	No
RAPGEF4	No
RIMS1	Maybe
RIMS2	Maybe
RYR1	Maybe
RYR2	Maybe
RYR3	No
APCS	No
COL5A1	Maybe
FBLN1	No
LAMA1	No
LGALS3BP	No
NID1	No
P4HA3	No

Gene symbol	Cancer involvement
P4HB	No
PLAU	No
PLG	No
TG	No
ANKRD1	No
MYBPC1	No
MYH2	Maybe
MYH9	Yes
OBSCN	No
TTN	Maybe
CADM1	No
CNTN2	No
CNTNAP2	Maybe
CNTNAP4	No
EPB41L3	Maybe
MACF1	Maybe
MYCBP	No
ROBO2	Maybe
SLIT1	No
SLIT2	Maybe
SLIT3	No
SSRP1	No

Gene symbol	Cancer involvement
AP4B1	No
APC	Yes
BUB1	No
PTPN13	No
ARFGEF2	No
GABRB1	No
GABRB2	Maybe
PIK3CA	Yes
EPHA2	No
EPHA3	Maybe
EPHA5	Maybe
EPHA6	No
LAMA4	Maybe
MAPK10	No
RRM2B	No
TP53	Yes

2.4 Discussion

SysSVM builds on previous efforts of our lab to identify novel cancer genes in individual tumours independently as opposed to focusing on recurrently altered genes across patient cohorts (D'Antonio and Ciccarelli 2013). This is useful in the context of precision medicine, as future therapeutics will be tailored to the mutational landscape of individual patients. In this chapter, I showed that systems-level properties can indeed distinguish known cancer genes from the rest of human genes. If used in a machine-learning framework, these properties can be utilised to identify novel driver genes. Such an approach requires no calculation of background mutational rate and, thus, it can reveal rare and patient-specific driver genes.

The main obstacle I faced while developing sysSVM was the definition of a negative set of observations, as currently there are no properties to define genes with no involvement in cancer. Soon it became apparent that a definition of an absolute set of non-cancer genes was challenging, but also irrelevant from a biological perspective, as several passenger genes in a certain cancer type might be drivers in others. In fact, the discovery of such genes is of great interest as context-specific driver potential of genes could highlight different underlying tumorigenic processes across cancer types.

Therefore, sysSVM models only the density of known driver genes (positive observations) in the input feature space and utilises one-class SVM for novelty detection (Schölkopf et al. 2001). SysSVM learns the decision boundary from the systems-level and molecular properties of known and cancer type-specific driver genes and, subsequently, scores all altered genes in each sample of a given

cohort according to their similarity to known drivers. The best models that were selected during the development phase of sysSVM showed high sensitivity (Table 2.4) and the construction of decision boundary utilised the majority of training observations. Overall, sysSVM operates as a meta-classifier and incorporates information from 4 different kernels (linear, radial, polynomial and sigmoid). This incorporation step was added because the high complexity of the input space and the multiple areas in which known cancer genes reside might not be sufficiently described by only one kernel. I showed that the sysSVM meta-score closely represented the raw predictions of individual kernels, as genes predicted by higher number of kernels had higher meta-score (Figure 2.10).

Assessment and ranking of each sysSVM feature revealed that on average molecular properties (sequencing data) were of lower importance for the classifiers than systems-level properties (Table 2.3 and Figure 2.9). However, the contribution of molecular properties towards predictions was not negligible, as their inclusion provided sysSVM with information on cancer type-specific alterations. For instance, the extensive genomic instability in OAC was captured by the high ranking of gene amplification in three out of four kernels used (Table 2.3).

As expected, predicted driver genes were mostly private to individual OACs (Table 2.5) and subject to genomic amplifications. Gene set enrichment analysis highlighted a number of enriched pathways with proven contribution to tumorigenesis, such as transcriptional regulation of *TP53* and DNA replication (Figure 2.11). The fact that sysSVM predictions, albeit not known driver genes, belong to well-known tumorigenic pathways is encouraging and denotes that high-scoring predictions might be true positives. Although experimental validation is invaluable to characterise the driver role of sysSVM predictions, I found literature

support for the tumorigenic role of multiple predicted drivers in the pilot cohort. For example, *AGO2*, a member of the RNA-induced silencing complex (RISC), was predicted in seven OACs (39%). *AGO2* has been reported to be involved in tumorigenesis through multiple miRNA-dependent pathways and has been found overexpressed in multiple carcinomas (Ye, Jin, and Qian 2015). Additionally, sysSVM predicted *KAT2B*, a histone acetyl-transferase, as a private top scoring prediction (Table 2.5). Closer examination of this gene revealed its involvement in the Notch signalling pathway, which has been previously reported to contribute to tumorigenesis in OAC (Wang et al. 2014), and its direct link with another top-scoring prediction, *E2F1*, which was predicted as a driver by sysSVM in three OACs (17%; Table 2.5). Pan-cancer expression analysis highlighted possible functional dependencies between *E2F1* and *KAT2B,* as well as its paralogue *KAT2A* (Figure 2.12). Finally, our first trial to experimentally validate *KAT2B* in the oesophageal cancer cell line OE19 confirmed its tumorigenic role in OAC (Figure 2.12).

In summary, sysSVM predictions in a pilot cohort of 18 OACs were particularly encouraging for further implementation of the method to larger cohorts, as they were in accordance with our initial hypothesis that rare or patient-specific cancer drivers can be predicted by systems-level and molecular properties. Although very limited, our initial experimental results and literature search supported the driver role of several predicted drivers. However, to fully explore the perturbed processes in OAC and identify novel genes and pathways involved in its tumorigenesis, a higher number of samples is needed. To this end, the third chapter of this thesis describes my first attempt to apply sysSVM to a cohort of 261 OACs from ICGC.

Chapter 3. Application of sysSVM to 261 Oesophageal adenocarcinomas

3.1 Chapter overview

In this chapter, I describe the application of sysSVM to a cohort of 261 OACs from ICGC to predict novel driver genes. The application of my method to a large cohort of samples allowed the optimisation of the algorithm and its refinement. Of particular interest was the fact that some of the OACs in this cohort had been published before (Secrier et al. 2016) and only a handful of drivers had been previously identified, suggesting that our understanding of the molecular landscape of this disease lags behind other cancer types. In the following paragraphs, I describe the landscape of rare and patient-specific sysSVM predictions, the perturbation of pathways to which these genes contribute to and their relevance to OAC. Interestingly, sysSVM predictions often converge towards the perturbation of similar biological pathways.

3.2 Introduction

Oesophageal cancer is the sixth leading cause of cancer-related mortality and the ninth most common cancer, affecting more than 550,000 people worldwide (Ferlay et al. 2015). The five-year survival rate ranges from 15% to 25% (Pennathur et al. 2013) and diagnosis at advanced stages is associated with poorer outcomes (Ek et al. 2013). The incidence of oesophageal cancer is rapidly increasing (Coleman, Xie, and Lagergren 2018; Thrift and Whiteman 2012). The

highest rates are observed along two geographic areas or “cancer belts”, one extending from China to Iran through central Asian countries and a second one including most Southeast African countries (Figure 3.1;Thrift, 2016).

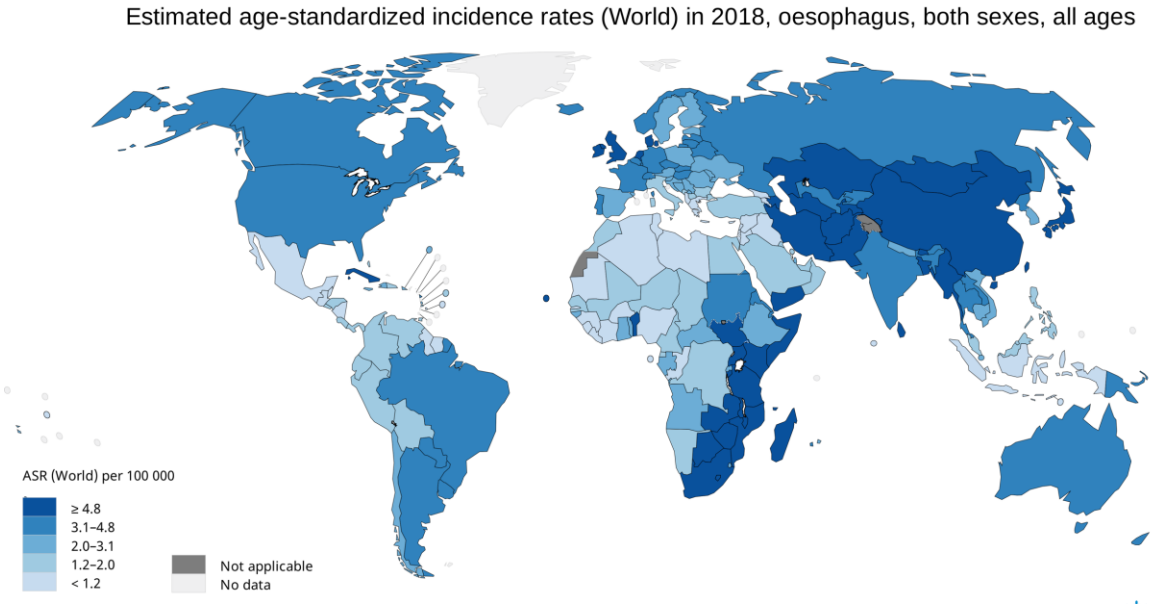


Figure 3.1. Incidence rate of oesophageal cancer. Data retrieved from GLOBOCAN 2018.

Histologically, oesophageal cancer is classified as oesophageal squamous cell carcinoma (OSCC) or oesophageal adenocarcinoma (OAC), but both types exhibit very distinct geographical distributions and unique clinical characteristics. OSCC is the most prevalent subtype, accounting for more than 80% of all cases of oesophageal cancer worldwide and shows a higher incidence rate in African American and Asian populations (Cook 2011; Dawsey et al. 1994). There are several factors that have been associated with increased risk of OSCC, including

tobacco use, alcohol consumption, low socioeconomic status, poor oral hygiene, and nutritional deficiencies (Pennathur et al. 2013).

During the last decades, however, the epidemiology of oesophageal cancer has changed in Western populations. There has been a rapid increase in the incidence of OAC, which has become the most predominant subtype in many developed countries (Coleman, Xie, and Lagergren 2018; Thrift and Whiteman 2012). OAC has surpassed OSCC in a number of industrialised countries and almost 50% of OAC cases worldwide are reported in Europe and North America. In particular, OAC incidence is high in United Kingdom, Ireland, France, Sweden and the Netherlands, indicating a bias towards Northern European populations (Rubenstein and Shaheen 2015). It is also worth noting that OAC exhibits a striking male predominance, as the incidence rates in all populations are significantly higher in men than women (Nordenstedt and El-Serag 2011; Cook, Chow, and Devesa 2009). The observed steep increase in OAC incidence cannot be explained by changes in the population's genetic make-up over a short timeframe of a few decades. Therefore, numerous epidemiological studies sought to investigate the contribution of non-genetic factors to OAC development. Based on the reported data, gastroesophageal reflux disease (GERD) (Lagergren et al. 1999), obesity (Thrift et al. 2014), and tobacco smoking (Cook et al. 2012, 2010) were identified as the major factors associated with increased risks of developing OAC. Taken together, these three factors account for nearly 80% of all cases of OAC in Western populations (Olsen et al. 2011).

GERD was identified as risk factor for OAC in the 1990s, when studies provided epidemiological evidence that GERD may predispose to oesophageal cancer (Chow et al. 1995; Lagergren et al. 1999; Farrow et al. 2000). A population-

based study conducted in Sweden reported that individuals with frequent (at least once weekly) symptoms of GERD (heartburn, regurgitation or both) had 8-fold higher probability of developing OAC, when compared to individuals without or with less frequent symptoms (Lagergren et al. 1999). Additionally, this risk was found further increased when the symptoms were persistent for over 20 years. These observations were further confirmed in subsequent independent studies (Pandeya, Olsen, and Whiteman 2013; Cook et al. 2014).

Although the molecular mechanism by which severe GERD may lead to OAC development is not yet fully elucidated, the prevailing hypothesis is that gastric acid induces chronic damage to the oesophageal lining. This may consequently lead to metaplastic changes, whereby the normal oesophageal squamous mucosa is replaced by a specialized columnar epithelium. This condition is called Barret's oesophagus (BE) and is considered a precursor of OAC (Reid et al. 2010). BE can indeed progress to low- and high-grade dysplasia, adenocarcinoma *in situ* and ultimately invasive adenocarcinoma (Anaparthy and Sharma 2014). The reported prevalence of BE is 1-2% in the general population and 10-15% in individuals with chronic reflux symptoms (Ronkainen et al. 2005; Voutilainen et al. 2000). As compared to the general population, patients with BE exhibit approximately a 10-fold higher risk of developing OAC (Hvid-Jensen et al. 2011) and are therefore periodically screened for dysplastic BE or early stage OAC.

Obesity has also been implicated in the development of OAC and it has been proposed that the increase in the incidence of OAC reflects the increasing obesity in Western populations (Kroep et al. 2014). However, the extent to which the increase in obesity can account for the rise in OAC has been debated (Kong et al. 2011). Obesity contributes to OAC development through promoting GERD. Excess

adipose tissue in the abdominal area leads to increased intra-gastric pressure and diminished lower oesophageal sphincter pressure, and subsequently promotes GERD (Friedenberg et al. 2008; Lagergren 2011). In addition, visceral adipose tissue can be metabolically active and release proinflammatory adipocytokines that can contribute to metaplastic and neoplastic changes (Nam et al. 2010). In support of this, increased serum levels of leptin and insulin are associated with increased risk of BE (Chandar et al. 2015). Moreover, the distribution of adipose tissue in the body may be more important than the overall adiposity when determining the risk of BE and OAC (Trevellin et al. 2015). This could partially explain the sex disparity observed in oesophageal adenocarcinoma (Singh et al. 2013; Steffen et al. 2015). Excess abdominal body fat, typically observed in males, is associated with increased risk for OAC (Steffen et al. 2015). On the contrary, gluteo-femoral adiposity, predominantly observed in women, is inversely correlated (Kendall et al. 2016).

Tobacco smoking has also been associated with increased risk for developing OAC. A pooled analysis of 12 studies conducted on OAC patients and control cohorts demonstrated that the risk of OAC is increased by two-fold among individuals that had a history of tobacco smoking (Cook et al. 2010). Furthermore, the same analysis reported a strong dose-response association between pack-years of smoking and OAC risk. Tobacco is a known carcinogen that has been linked to higher DNA damage in Barrett's mucosa (Olliver et al. 2005). Another potential mechanism through which cigarette smoking might promote OAC is its ability to relax the lower oesophageal sphincter, therefore predisposing to GERD (Kadakia, De la Baume, and Shaffer 1996).

Curative treatment in OAC is currently based on esophagectomy, with additional peri-operative chemotherapy or chemoradiotherapy (Allum et al. 2009; Cunningham et al. 2006; van Hagen et al. 2012). In contrast to other cancer types, the use of targeted agents has been lagging behind and many recent phase III trials reported disappointing or inconclusive results (Woo, Cohen, and Grim 2015; Young and Chau 2016; Kopp and Hofheinz 2016). To date, only trastuzumab, a monoclonal antibody targeting *HER2*, has led to an improvement in patient outcome (Bang et al. 2010). Advances in this area have been hindered by the fact that OAC is characterised by extensive genomic instability, manifested in both somatic mutational burden and copy number variations (The Cancer Genome Atlas Research Network 2017). As a result, the number of potential tumorigenic alterations that may confer selective advantage to cancer cells, and, thus drive their growth is high. Therefore, the identification of genes that play an integral part in this process, i.e. driver genes, is critical to fully understand the molecular determinants of OAC and to inform the development of targeted therapeutic approaches.

OAC, as a genomically unstable cancer type, is characterized by widespread inter-patient heterogeneity, with very few alterations exceeding 10% of recurrence across patients (Secrier et al. 2016). This makes the full characterisation of driver events particularly challenging, as the genomic landscape of OAC is highly variable and recurrent events are rare. Therefore, current methods, which are based on the recurrence of gene alterations to identify drivers are insufficient. OAC genomes are characterised by complex patterns of rearrangements and genomic catastrophes (Nones et al. 2014). Chromosomal shattering (chromothripsis) and mis-segregation of chromosomes during cell cycle can affect several genes, as opposed to single

nucleotide variants, and have been shown to drive cancer development (Zhang, Leibowitz, and Pellman 2013). For example, chromothripsis might contribute to the generation of double-minute chromosomes, which are extrachromosomal elements containing multiple copies of genes, whose amplification can lead to an oncogenic phenotype (Sanborn et al. 2013). Well-known drivers in OAC are genes involved in tumorigenesis of multiple other cancer types, such as *TP53*, *CDKN2A*, *SMARCA4*, *ARID1A, SMAD4*, *ERBB2* and amplifications of *VEGFA, ERBB2, EGFR, GATA4/6, CCNE1* (Agrawal et al. 2012; Dulak et al. 2013; Weaver et al. 2014; Ross-Innes et al. 2015; Secrier et al. 2016; The Cancer Genome Atlas Research Network 2017). The low number of driver genes leaves approximately 10-20% of patients without known genetic determinants and often the number of identified drivers per sample is too low to fully explain the disease.

OAC provides a good example of a cancer type where genes altered in low frequencies could play a patient-specific tumorigenic role, and as such, it is ideal for the development of new methods aiming to identify rare cancer drivers. I hypothesised that alongside the critical role of recurrent and well-known cancer drivers, complementary somatic alterations of several other genes help cancer progression in individual patients. These “cancer helper” genes have potentially a very broad spectrum of functions, ranging from optimisation of the function of major drivers to the promotion of competition between or even cooperation of cancer cells (explained in chapter 1).

To comprehensively characterise the molecular mechanisms relevant to OAC, I applied sysSVM to 261 OACs from the UK OCCAMS Consortium. I first trained the classifier using 34 properties specific to known cancer genes (31 of

which were described in chapter 2) and then prioritized 952 genes that, together with the known drivers, promote cancer development. By applying sysSVM to a much larger cohort than the one used in the pilot phase of this thesis (chapter 2), apart from the characterisation of 'helper' genes relevant to OAC, I had the opportunity to further optimise several parts of my method. These are described in the following paragraphs and include: i) the assessment of convergence of the best parameters during cross validation, ii) the usage of additional features, such as structural variations and iii) the validation of the trained models to additional cohorts.

3.3 Results

3.3.1 OAC mutational landscape

WGS data from a diverse collection of OACs, including Siewert types 1, 2 and 3 (Table 3.1), allowed me to examine the types of alterations that dominate the mutational landscape of this disease. Confirming previous studies (Dulak et al. 2013), OACs in our cohort had a high burden of point mutations with an average of 166 protein-coding SNVs per sample (ranging from 1 to 860) (Figure 3.2A). However, OAC mutation landscape was dominated by large-scale alterations, such as genomic amplifications and structural variations. On average, copy number and structural somatic aberrations affected twice as much protein-coding genes as compared to point mutations (384 genes per sample ranging from 17 to 1,840) (Figure 3.2A). Most OACs in our cohort exhibited a highly fragmented genomic landscape with large chromosomal regions being subject to genomic amplification

or loss (Figure 3.2B). Overall, the number of OAC genomes undergoing catastrophic chromosomal events is higher than those in other cancer types (Stephens et al. 2011; Molenaar et al. 2012). Unlike melanoma or pancreatic adenocarcinoma, which are driven by oncogenic point mutations in *KRAS* and *BRAF*, OAC is driven by oncogenic amplifications (Nones et al. 2014).

In terms of point mutations, known driver genes in OAC were highly heterogeneous amongst the cases included in our cohort. With the exception of *TP53*, which was found mutated in more than 70% of OACs, all other known drivers were mutated in less than 15% of OACs (Figure 3.3). Taken together, these observations confirmed that the mutational landscape of OAC is highly heterogeneous and is dominated by large-scale genomic catastrophes. As mentioned in the introduction of this chapter, therapeutic approaches using targeted therapies in OAC lag behind those of other cancer types and this is due to the limited understanding of cancer drivers in this cancer type. Given that one of the main attributes of sysSVM is its ability to integrate both point mutations and structural variations to identify cancer drivers, I considered OAC and its heterogeneous mutational landscape the ideal disease to apply my method to.

Table 3.1. Summary of clinical characteristics of the 261 OAC samples. Percentages that do not sum up to 100% within each category denote missing data. Data were collected from the ICGC. Gastro-oesophageal junction (GOJ) types are reported according to Siewert classification.

Mean age at diagnosis (stDev)	65.5 (9.97)
Sex (% male)	84.7
Treatment (%)	
Chemotherapy	28.7
No treatment	47.1
Radiation therapy	0.3
Surgery	0.6
Other therapy	0.3
Tumour location (%)	
GOJ Type 1	34.9
GOJ Type 2	25.3
GOJ Type 3	6.1
Primary tumour (%)	
Stage I	18.8
Stage II	14.2
Stage III	43.7
Stage Iv	2.3
Node positive (%)	50.9
Metastasis (%)	11.1

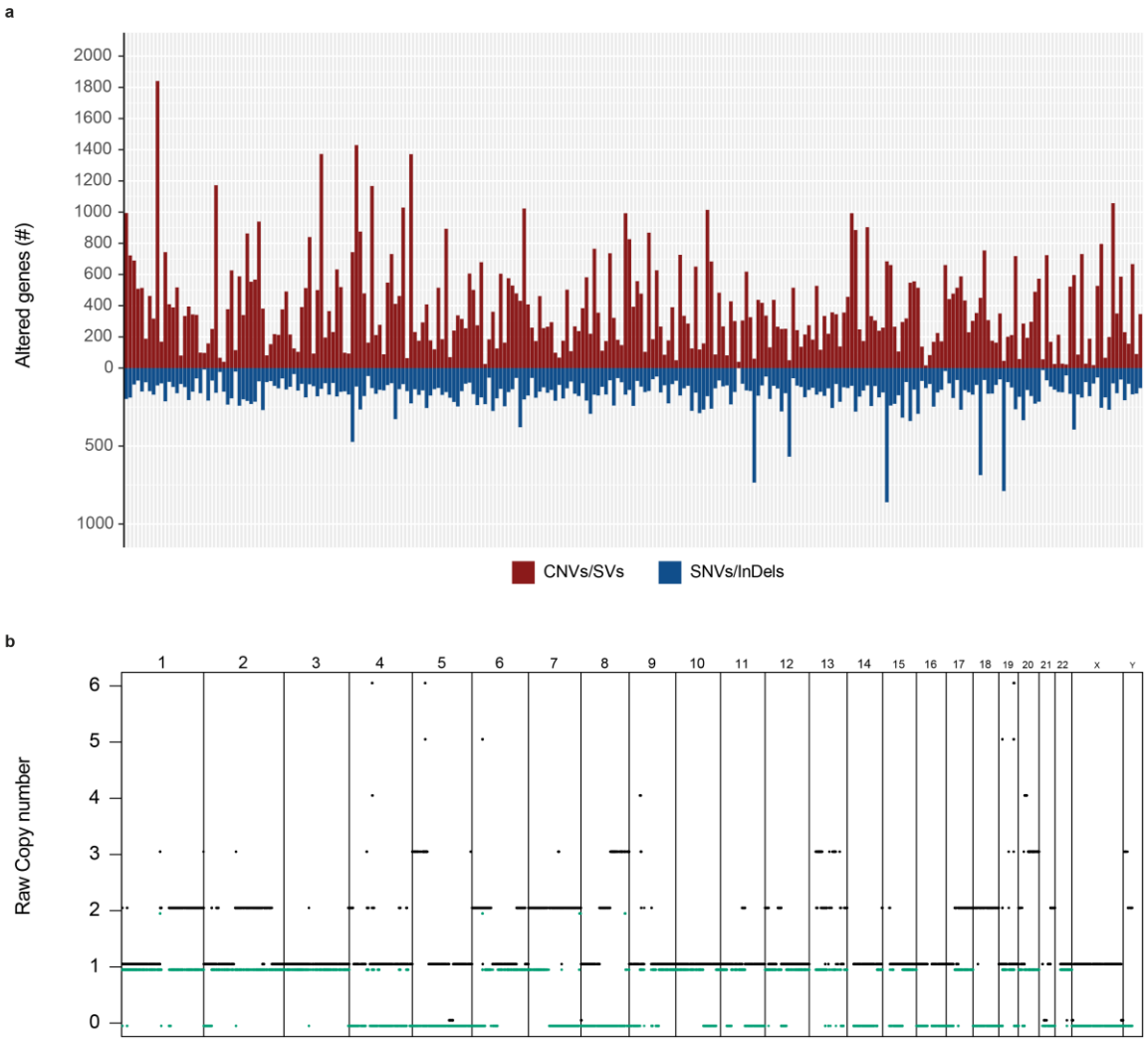


Figure 3.2. Overview of genomic alterations in the OAC cohort (n=261). **(a)** The total number of protein-coding genes affected by copy number alterations or structural variants (above x axis) and point mutations or indels (below x axis). **(b)** A representative copy number profile of an OAC tumour (LP6008031-DNA_A01) affected by large-scale amplifications and deletions. The two colours (green and black) represent the copy number of major and minor alleles (y axis) across different chromosomes of the human genome (x axis).

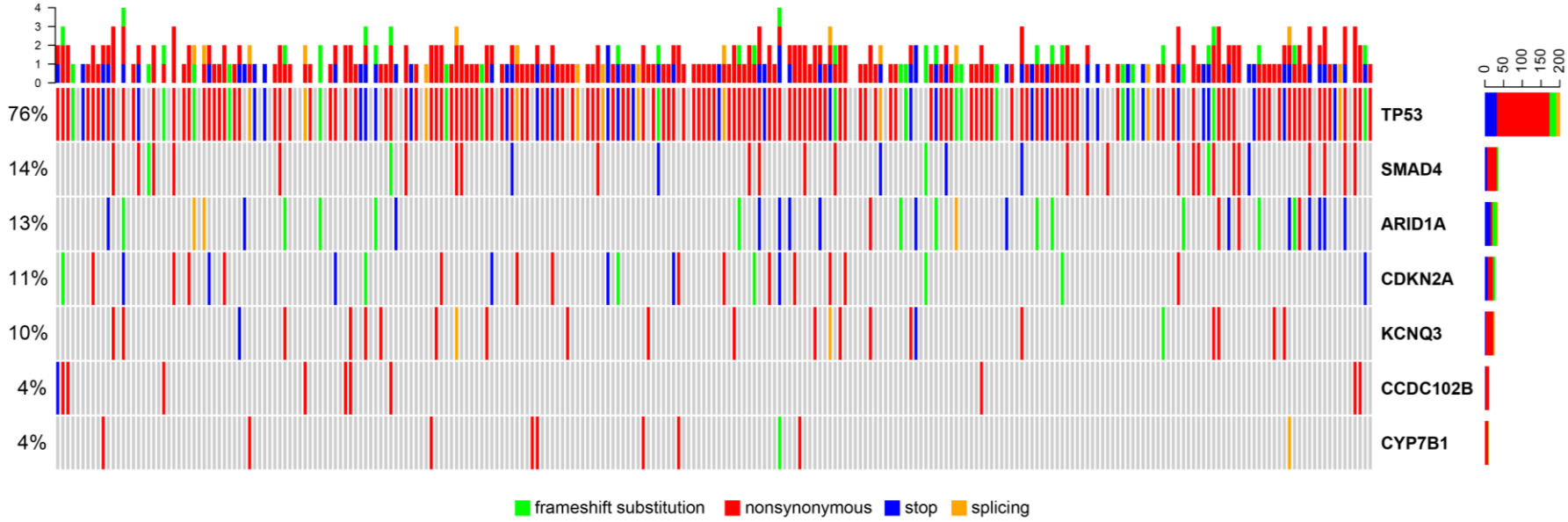


Figure 3.3. Frequency of frameshift, nonsynonymous, stop gain or loss and splicing alterations for 7 seven driver genes that have been previously identified in OAC (Secrier et al. 2016)**.**

3.3.2 SysSVM workflow

SysSVM applies supervised machine learning to predict altered genes contributing to cancer based on the similarity of their systems-level and molecular properties to those of known cancer genes (as mentioned in chapter 2). Systems-level properties are genomic, epigenomic, evolutionary and gene expression features that distinguish cancer genes from the rest of human genes and were described in detail in the previous chapter. Briefly, they include gene length and protein domain organisation (D'Antonio and Ciccarelli 2013; An et al. 2016), gene duplicability (Rambaldi et al. 2008; D'Antonio and Ciccarelli 2011), chromatin state (Lieberman-Aiden et al. 2009), connections and position of the encoded proteins in the protein-protein interaction network (Rambaldi et al. 2008), number of associated regulatory miRNAs (D'Antonio and Ciccarelli 2011), gene evolutionary origin (D'Antonio and Ciccarelli 2011) and breadth of gene expression in human tissues (D'Antonio and Ciccarelli 2013; An et al. 2016). Moreover, as more sequencing data became available from ICGC, I extended the catalogue of the molecular properties from seven to 10, including data on structural variation (Table 3.2). In particular, I added translocations, inversions, and insertions to the molecular features of sysSVM. Also, to control for the high number of amplifications in OAC, I corrected gene copy number, using the estimated ploidy of each sample (see Methods). In addition to what was described in chapter 2, I added an extra step to the sysSVM algorithm in order to check for the convergence of cross-validation (see step 3 below).

In its optimised version, sysSVM workflow is composed of three main conceptual steps (Figure 3.4A):

- **Step 1: Feature mapping.** sysSVM maps 34 features of all altered genes in the sample cohort under study. Ten of the 34 features are

derived from molecular properties and 24 features are derived from systems-level properties of known cancer genes, as summarized in Table 3.2. These 34 features are used to define the regions of the feature space where the known cancer genes reside. Twenty-two of them are categorical and 12 are continuous or discrete variables.

- **Step 2: Model selection.** sysSVM applies a grid search to optimise the parameters used in each kernel. The parameter space in which sysSVM searches was described in detail in chapter 2. Briefly, 18 combinations of parameters are examined for the linear kernel, 216 combinations for the radial and sigmoid kernels and 648 combinations for the polynomial kernel, for a total of 1,098 parameter combinations. To identify the best combination of parameters for each kernel, a user-defined number of iterations of three-fold cross-validation is performed (default = 10,000). At each iteration, the genes of the training set are randomly split into two subsets, one used for training (2/3 of the genes) and one as a test set (1/3 of the genes). Predictions are performed on the test set and the sensitivity of each set of parameters is computed, as described in chapter 2. To ensure robustness of cross-validation and to avoid artificial inflation of sensitivity, because the same gene can be altered in multiple samples, the split of training and test set is performed on the unique set of genes. This results in slight variations of the size of training/test sets during the cross-validation. Subsequently, the distribution of sensitivity every *n* iterations (default = 100) is derived. The least variant model among the top

five most sensitive models in each kernel (considering the mean sensitivity) is chosen as the best model for that kernel, as described in chapter 2. To account for the effect of increasing number of cross validation iterations, at each increment of *n* cross validations (default = 100), the selection of best models takes into account all previous cross-validation iterations. To account for the effect of the order of iterations, this cumulative assessment is repeated a number of times (default = 5) where the iterations of cross validation are randomly reordered. This produces *m* sets of best models (from a default of 5 re-orderings of 100 increments, *m* = 500).

- **Step 3: Training and prediction.** All *m* sets of best models identified in step 2 are used for training using the whole training set. This is a modification of sysSVM compared to what was described in chapter 2, as we[2] sought to investigate how the number of cross-validation iterations affected the selection of the best models. For each set of best models, cancer genes are predicted in individual samples and combined using the formula 2.10. This results in *m* lists of top *k* genes (default *k* = 10) in each sample. The most frequent list of top *k* genes overall is selected as the final list of predicted cancer genes.

[2] Analysis was performed by Damjan Temelkovski

I applied sysSVM to 261 OACs from OCCAMS and the ICGC, which are summarised in Table 3.3. In the first step, I extracted 17,078 genes with (i) truncating, (ii) non-truncating damaging or (iii) gain-of-function mutations, (iv) homozygous deletions, (v) amplifications, (vi) insertions, (vii) inversions and (vii) translocations in the whole cohort (median of 382 damaged genes per patient, Figure 3.4B, Table 3.3) and mapped their 34 molecular and systems-level features. As expected, most of the altered genes were amplified (average of 307.9 amplified genes per sample; Table 3.3), owing to the high genomic instability in OAC. In the second step, 476 of the extracted genes, harbouring 4,091 damaging alterations in total, were annotated as known cancer genes based on data provided from Cancer Gene Census (Forbes et al. 2017) and are listed in appendix table 7.1. These known cancer genes comprised the training set of sysSVM. To optimise the parameters of each kernel, three-fold cross-validation was ran with 10,000 iterations and the results were combined every 100 iterations to obtain 500 sets of best parameters for the four kernels (Table 3.4), as described in step 3 above. In the third step, the best models from step two (one per kernel each 100 iterations) were used for training and prediction. All genes except those used for training (n=16,602) were first scored in each patient individually by combining the decision values of the four kernels (using the formula 2.10) and then ranked according to the resulting score. Overall, the best parameters in each kernel converged on a limited number of values during cross-validation (Table 3.4). Multiple sets of parameters could result in the same set of genes and a single set of parameters could result in slightly different sets of genes (usually differing by one or two genes) depending on the cross-validation sensitivity. In order to address the fact that in some cases slightly different parameters were equally good and because I was interested in

the list of top 10 scoring genes that were most frequently predicted in each sample, I selected 952 genes (referred to as 952A in Table 3.4) for downstream analysis. Overall, these genes along with the almost identical gene lists 952B, 952C, 951A, 951B, 950 accounted for 434 out of the 500 gene lists (86.8%; Table 3.4).

Taken together, these results highlight a set of 952 genes that were predicted by sysSVM as drivers of tumorigenesis in OAC. Moreover, the application of sysSVM to a larger cohort than the one used for its development (18 OACs described in chapter 2) allowed me to refine the algorithm, assess the robustness of selection of the best parameters and thereby ensure reproducibility and stability of the predicted genes.

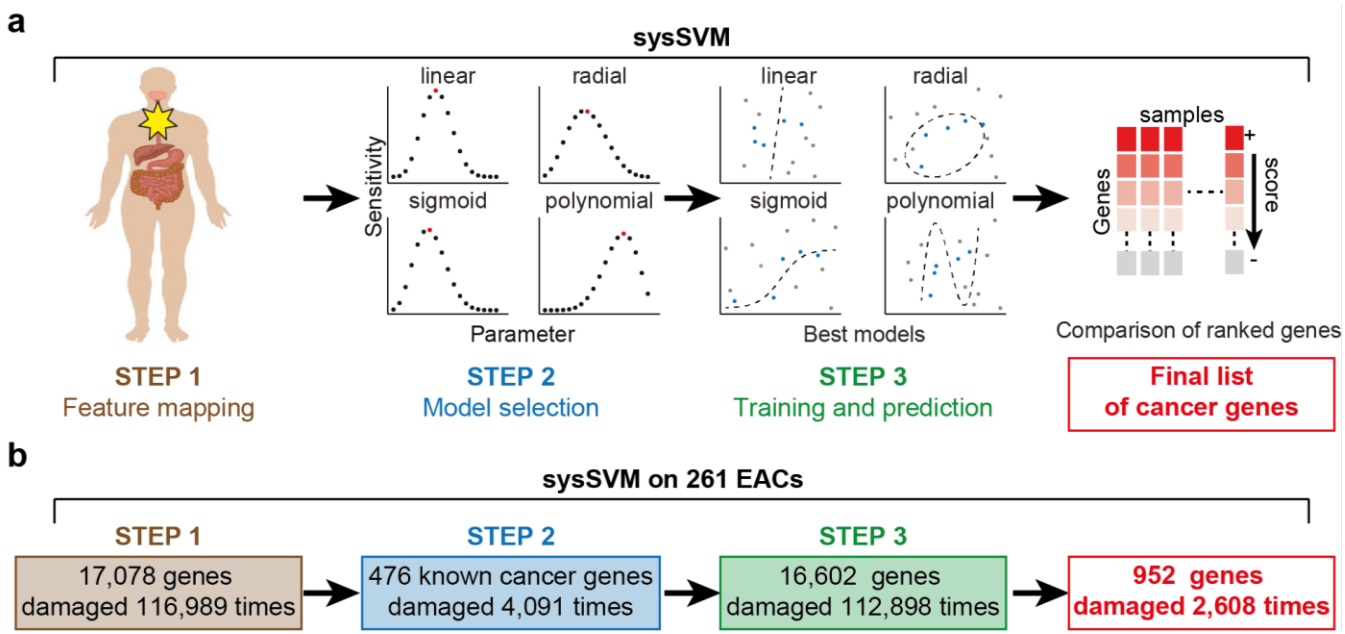


Figure 3.4. Overview of sysSVM. **(a)** Step-by-step description of sysSVM, which was used to identify cancer genes on 261 OACs. **(b)** Genes extracted for each sysSVM step. Genes with somatic damaging alterations (n=116,989) were extracted from 261 OACs and divided into training (known cancer genes, blue) and prediction (rest of altered genes, green) sets. All altered genes (n=112,898) were scored in each patient individually after the selection of the best set of parameters per kernel during cross-validation. The final list of top-scoring genes which was used for downstream analysis was comprised of 952 genes.

Table 3.2. Description of sysSVM features. Listed are 10 molecular and 24 systems-level features used in sysSVM. For each of them, described are: the original gene property, whether it is categorical or continuous, its operational definition (see Methods) and the number of unique and redundant (in brackets) genes in 261 OACs. The description of systems-level properties of cancer genes is also given. For all systems-level properties, except gene length, duplication status and ohnologs, the number of unique genes before imputation is given (see Methods). CG = cancer gene; TSG = tumour suppressor gene; OG = oncogene; WGD = whole genome duplication; n = number; l = length.

Gene property	Feature for classification	Type	Cancer gene feature	Operational definition	Genes in 261 OACs
Copy number variation	Gene gain	Molecular (categorical)	Not applicable	CN >= 2*sample ploidy	13,622 (79,216)
	Gene loss	Molecular (categorical)	Not applicable	CN = 0	1,117 (3,089)
	Gene copy number (n)	Molecular (continuous)	Not applicable	Somatic copy number (ASCAT)	17,078 (116,989)
Structural variation	Gene translocation	Molecular (categorical)	Not applicable	Somatic translocation event (Manta)	5,577 (11,137)
	Gene inversion	Molecular (categorical)	Not applicable	Somatic inversion event (Manta)	5,546 (10,320)
	Gene insertion	Molecular (categorical)	Not applicable	Somatic insertion event (Manta)	519 (646)
SNVs and indels	Truncating alterations (n)	Molecular (continuous)	Not applicable	Stopgain, stoploss, frameshift alterations (ANNOVAR)	1,992 (2,471)
	Non-truncating damaging alterations (n)	Molecular (continuous)	Not applicable	Damaging nonframeshit, nonsynonymous, splicing alterations (dbNSFP)	7,287 (15,508)
	Gain of function alterations (n)	Molecular (continuous)	Not applicable	Gain of function (OncodriveClust)	170 (614)
	All exonic SNVs and indels (n)	Molecular (continuous)	Not applicable	Silent and non-silent alterations (ANNOVAR)	8,359 (18,941)
Gene length	Gene length (l)	Systems-level (continuous)	CGs tend to be long (1)	Length of the longest isoform (RefSeq)	17,078
Gene duplication status	Gene duplication status	Systems-level (categorical)	TSGs are enriched in single-copy genes (2)	Mapping on >1 gene locus for ≥60% of protein length	17,078
Whole genome duplication	Ohnolog	Systems-level (categorical)	OGs are enriched in ohnologs (3)	Gene duplicate retained after whole genome duplications	17,078
Protein domains	Protein domains (n)	Systems-level (continuous)	CGs are enriched in multi-domain proteins (4)	Number of protein domains (CDD)	17,039
Chromatin state	Chromatin state	Systems-level (continuous)	CGs localise preferentially in open chromatin (5)	Chromatin state from Hi-C experiment in K562 cells.	14,959
Protein-protein interaction network	Protein degree (n)	Systems-level (continuous)	CGs encode preferentially protein hubs (2)	Number of connections in the protein-protein interaction network	13,268
	Hub	Systems-level (categorical)		Top 25% most connected proteins	
	Protein betweenness (n)	Systems-level (continuous)	CGs encode preferentially central proteins (2)	Centrality in the protein-protein interaction network	

	Central protein	Systems-level (categorical)		Top 25% most central proteins	
miRNA interaction network	miRNA interactions (n)	Systems-level (continuous)	CGs tend to be regulated by a larger number of miRNAs (3)	Number of miRNAs interacting with the gene	10,689
Evolutionary origin	Old gene	Systems-level (categorical)	TSGs are enriched in old genes and OGs are enriched in genes originated in Metazoans (3)	The gene originated before metazoans	16,354
	Origin in prokaryotes	Systems-level (categorical)		Oldest ortholog found in prokaryotes	
	Origin in single cell eukaryotes	Systems-level (categorical)		Oldest ortholog found in eukaryotes	
	Origin in opisthokonts	Systems-level (categorical)		Oldest ortholog found in opisthokonts	
	Origin in metazoans	Systems-level (categorical)		Oldest ortholog found in metazoans	
	Origin in vertebrates	Systems-level (categorical)		Oldest ortholog found in vertebrates	
	Origin in mammals	Systems-level (categorical)		Oldest ortholog found in mammals	
	Origin in primates	Systems-level (categorical)		Oldest ortholog found in primates	
Expression	Ubiquitously expressed	Systems-level (categorical)	CGs are enriched in genes ubiquitously expressed (1,4)	Gene is expressed in > 28/30 tissues	16,728
	Medium expressed	Systems-level (categorical)		Gene is expressed in 3-28 tissues	
	Selectively expressed	Systems-level (categorical)		Gene is expressed in 2-3 tissues	
	Specifically expressed	Systems-level (categorical)		Gene is expressed in 1 tissue	
	Not expressed	Systems-level (categorical)		Gene is expressed in 0 tissues	
	Tissues where the gene is expressed (n)	Systems-level (continuous)		Number of tissues	

Table 3.3. Description of somatically altered genes in the 261 OACs. For each sample, the number of genes with predicted damaging alterations is provided. The total number of altered genes is 116,989 (17,087 unique hits).

Sample	Genes with damaging alterations (n)								
	Gain	Loss	Trans location	Inversion	Insertion	Truncating	Non-truncating damaging	Gain of function	Total
LP6005334-DNA_A02	330	23	26	29	2	8	49	0	442
LP6005334-DNA_A03	506	0	9	16	0	17	57	1	589
LP6005334-DNA_B01	285	1	29	32	0	5	36	3	373
LP6005334-DNA_C01	241	17	44	47	0	10	49	1	393
LP6005334-DNA_C03	35	2	12	39	0	5	57	1	148
LP6005334-DNA_D01	86	0	16	16	0	1	2	1	99
LP6005334-DNA_D03	199	32	15	13	0	1	34	0	279
LP6005334-DNA_H02	494	0	45	32	0	9	79	4	638
LP6005409-DNA_A02	857	23	43	33	0	4	26	1	965
LP6005409-DNA_E02	74	41	13	12	0	6	38	0	180
LP6005409-DNA_F01	109	9	45	74	1	2	43	1	262
LP6005500-DNA_A02	476	0	12	36	0	12	72	4	590
LP6005500-DNA_B02	405	16	37	53	2	10	70	1	580
LP6005500-DNA_C01	159	21	4	11	0	4	33	1	229
LP6005500-DNA_D01	190	1	23	18	0	2	41	1	269
LP6005500-DNA_H03	53	0	19	18	2	10	48	6	149
LP6005690-DNA_A02	473	21	33	30	2	5	47	0	598
LP6005690-DNA_B03	420	1	21	29	0	10	43	0	512
LP6005690-DNA_C01	899	22	103	22	3	2	37	1	1063
LP6005690-DNA_D03	250	3	27	16	1	8	51	3	350
LP6005690-DNA_F01	131	6	21	26	3	10	73	7	258
LP6005690-DNA_G01	91	38	30	45	1	9	71	2	265
LP6005690-DNA_H01	27	1	28	19	2	11	64	2	142
LP6005935-DNA_B01	516	9	35	85	0	6	33	2	636
LP6005935-DNA_C03	171	0	6	10	0	6	21	1	211
LP6005935-DNA_H02	400	3	43	52	1	12	59	2	530
LP6007401-DNA_A01	64	0	28	25	0	5	75	5	178
LP6007409-DNA_A01	418	3	61	27	1	8	59	1	566
LP6007436-DNA_A01	134	5	14	16	0	7	64	5	230
LP6007512-DNA_A01	952	6	19	36	2	12	65	5	1062
LP6007538-DNA_A01	248	3	11	5	1	10	57	5	331
LP6008031-DNA_B01	422	0	30	52	0	5	31	1	518
LP6008031-DNA_B03	52	20	9	7	0	7	42	2	130
LP6008031-DNA_C01	371	17	23	33	3	13	86	2	524
LP6008031-DNA_D02	573	0	32	32	3	7	56	2	676
LP6008031-DNA_E01	38	0	14	0	4	60	281	6	393
LP6008031-DNA_H02	10	4	18	10	0	36	192	2	261
LP6008051-DNA_A02	169	23	24	42	2	8	40	0	287
LP6008051-DNA_B01	74	1	37	33	2	3	54	3	191
LP6008051-DNA_C02	45	45	49	40	13	8	44	1	234
LP6008141-DNA_B01	241	51	41	19	0	7	47	1	393
LP6008141-DNA_B02	295	3	102	68	11	6	41	3	500
LP6008141-DNA_D01	649	17	157	129	10	25	98	0	995
LP6008202-DNA_A01	512	36	106	67	10	10	139	11	829
LP6008202-DNA_B01	187	1	51	28	7	11	83	3	350
LP6008202-DNA_B02	49	11	26	34	0	6	65	2	176
LP6008202-DNA_C02	277	0	25	31	0	7	30	4	347
LP6008202-DNA_D01	490	7	10	37	1	5	40	7	584
LP6008202-DNA_E01	441	0	53	50	3	12	121	7	644
LP6008202-DNA_E02	82	2	41	12	3	7	32	0	170
LP6008221-DNA_A02	86	10	53	42	3	6	60	0	236
LP6008221-DNA_C02	465	20	16	22	0	6	23	1	545
LP6008221-DNA_F02	359	0	44	70	4	115	226	1	780
LP6008280-DNA_C01	630	28	38	61	1	16	95	2	822
LP6008280-DNA_E02	398	16	56	43	8	8	89	3	579
SS6003109	36	1	26	25	1	9	67	1	160
SS6003115	154	5	21	14	0	8	80	1	270
SS6003308	326	1	29	249	10	6	29	3	621
LP2000104-DNA_A01	963	3	16	26	0	13	74	4	1076
LP6005334-DNA_E01	1002	3	133	97	1	8	62	3	1235
LP6005334-DNA_F01	316	2	44	27	2	18	72	2	466
LP6005334-DNA_H01	775	9	32	76	2	16	71	1	940
LP6005500-DNA_A01	305	1	61	40	1	7	34	2	433
LP6005500-DNA_F02	817	3	43	16	0	10	98	2	967
LP6005500-DNA_G03	1108	3	32	46	2	8	39	0	1208
LP6005690-DNA_C03	1323	4	17	32	4	9	90	6	1465
LP6005935-DNA_B02	459	2	24	35	0	8	76	1	582
LP6005935-DNA_C04	111	4	235	29	1	11	97	6	471
LP6005935-DNA_E02	416	2	86	110	0	9	49	1	632

LP6005935-DNA_E03	362	86	61	43	1	6	54	1	585
LP6005935-DNA_G01	960	0	37	68	2	6	74	2	1096
LP6005935-DNA_G02	294	43	40	63	0	9	65	1	475
LP6007430-DNA_A01	708	0	25	49	0	9	76	2	847
LP6007523-DNA_A01	453	1	13	19	2	4	45	3	525
LP6007531-DNA_A01	785	19	42	51	0	7	62	8	937
LP6007544-DNA_A01	351	18	8	15	0	9	50	0	447
LP6007550-DNA_A01	701	1	10	15	0	9	74	4	797
LP6008031-DNA_G01	354	20	52	12	4	10	42	1	486
LP6008141-DNA_E01	174	16	22	48	2	15	61	1	321
LP6008221-DNA_B02	368	1	45	46	1	5	42	4	487
SS6003111	703	0	25	19	0	14	64	3	802
SS6003121	715	1	53	40	0	11	101	4	900
LP6005334-DNA_G02	275	0	44	33	1	8	69	6	414
LP6005409-DNA_G03	452	1	20	18	0	8	55	1	542
LP6005409-DNA_H01	167	1	17	46	0	11	45	0	267
LP6005500-DNA_E01	0	25	61	13	5	14	56	0	164
LP6005935-DNA_E01	97	12	26	39	2	13	89	3	255
LP6007404-DNA_A01	35	0	8	30	0	9	42	1	118
LP6007440-DNA_A01	287	4	28	21	0	10	80	6	410
LP6007520-DNA_A01	480	0	49	67	0	14	30	0	601
LP6008031-DNA_D01	192	43	39	41	1	5	35	1	330
LP6008031-DNA_E02	386	5	25	57	8	12	54	3	503
LP6008051-DNA_E02	243	24	43	28	9	11	65	4	408
LP6008141-DNA_C01	913	1	51	58	3	9	44	2	1043
LP6008221-DNA_D01	443	39	101	62	3	15	102	3	695
LP6008280-DNA_B01	19	2	5	8	0	48	290	9	374
LP6008280-DNA_F01	505	18	25	47	0	10	77	1	653
LP6008280-DNA_H01	119	17	21	31	1	12	42	0	222
SS6003149	114	6	62	30	0	13	91	5	301
SS6003314	131	0	6	27	0	3	38	3	196
LP2000106-DNA_A01	577	50	63	57	1	5	40	15	730
LP2000108-DNA_A01	354	52	34	65	0	2	38	2	530
LP2000328-DNA_A01	306	45	55	62	3	13	46	2	509
LP2000332-DNA_A01	654	7	68	24	4	14	76	3	827
LP6005334-DNA_A04	2	17	39	36	0	8	24	1	112
LP6005334-DNA_E03	26	1	8	4	0	9	53	1	100
LP6005409-DNA_D02	344	2	13	33	0	11	97	3	483
LP6005409-DNA_E01	17	28	16	20	0	6	36	2	118
LP6005409-DNA_F03	317	5	51	34	0	5	27	0	406
LP6005500-DNA_E03	487	14	175	99	3	19	188	9	923
LP6005690-DNA_C02	2	11	43	14	3	9	61	1	131
LP6005690-DNA_E01	270	24	92	36	2	19	86	3	507
LP6005690-DNA_H02	188	22	19	13	2	18	77	6	336
LP6005935-DNA_G03	222	0	33	35	0	5	22	0	283
LP6007420-DNA_A01	127	7	78	36	0	7	61	2	301
LP6007434-DNA_A01	65	16	21	14	1	9	53	3	173
LP6007542-DNA_A01	104	20	30	26	0	10	93	3	278
LP6007552-DNA_A01	266	13	34	41	0	10	53	1	396
LP6007602	131	4	17	11	3	9	98	4	263
LP6008031-DNA_A01	752	181	31	55	3	7	74	2	1073
LP6008031-DNA_A02	30	1	496	125	123	14	106	2	786
LP6008031-DNA_D03	260	1	62	32	0	10	45	1	375
LP6008031-DNA_F01	355	1	34	48	1	3	45	1	464
LP6008031-DNA_F02	281	16	22	23	0	2	17	0	350
LP6008031-DNA_G02	216	1	24	30	4	13	37	0	314
LP6008051-DNA_A01	456	2	53	56	1	4	30	0	536
LP6008051-DNA_D01	454	3	31	60	0	12	69	1	600
LP6008141-DNA_A01	74	16	35	32	5	14	78	2	231
LP6008141-DNA_F02	261	3	45	35	2	13	93	4	431
LP6008221-DNA_B01	638	0	15	12	1	0	8	1	666
LP6008221-DNA_D02	393	1	28	17	2	6	40	1	478
LP6008221-DNA_E01	82	6	112	49	8	7	46	3	281
LP6008221-DNA_G01	665	24	47	59	0	7	22	0	777
LP6008280-DNA_B03	177	2	9	19	6	14	34	1	258
LP6008280-DNA_D02	8	1	132	60	19	4	52	2	247
LP6008280-DNA_G02	637	49	20	33	1	5	22	0	746
SS6003317	575	40	43	31	1	8	61	1	732
LP6005334-DNA_D02	21	15	60	28	0	8	81	1	207
LP6005409-DNA_F02	167	17	16	23	0	9	46	1	268
LP6005409-DNA_H03	79	0	22	31	0	2	15	2	141
LP6005500-DNA_C03	321	28	9	11	0	10	56	1	428
LP6005690-DNA_D01	156	20	32	30	0	14	43	1	288
LP6005690-DNA_D02	56	0	91	30	1	7	49	4	225
LP6005690-DNA_F02	40	0	43	48	0	6	60	4	186
LP6005935-DNA_A03	202	3	44	48	5	15	36	0	298
LP6005935-DNA_C02	0	6	10	8	0	13	88	4	125
LP6007398-DNA_A01	169	24	88	28	1	11	52	1	355
LP6007416-DNA_A01	150	0	73	75	1	9	66	1	338
LP6007508-DNA_A01	110	1	30	13	1	3	29	1	179

LP6007516-DNA_A01	734	30	36	40	0	3	48	1	873
LP6007567-DNA_A01	171	6	91	50	1	7	49	3	338
LP6007597	64	0	30	26	0	11	110	2	235
LP6008031-DNA_A04	13	1	57	16	0	5	50	3	139
LP6008031-DNA_B02	180	11	44	43	1	11	42	1	316
LP6008051-DNA_G01	220	0	72	70	0	8	41	1	391
LP6008141-DNA_F01	785	0	70	101	3	10	37	1	945
LP6008141-DNA_G01	163	85	24	52	0	3	50	4	355
LP6008141-DNA_G02	95	67	49	36	7	8	66	2	313
LP6008202-DNA_G01	40	5	15	25	0	4	27	1	110
LP6008221-DNA_A01	138	15	51	55	2	6	54	3	279
LP6008280-DNA_A01	119	5	30	12	1	6	39	1	208
SM-4AX85	156	20	27	19	0	5	50	2	265
SM-4AX86	0	0	17	11	0	6	62	3	96
SM-4AX87	0	0	10	15	0	1	20	0	43
SM-4B295	478	4	27	18	0	7	54	3	582
SS6003302	227	5	39	832	10	6	34	1	1092
LP2000105-DNA_A01	584	23	105	64	9	11	61	16	801
LP2000325-DNA_A01	347	14	119	51	3	9	55	6	564
LP2000327-DNA_A01	137	17	6	30	0	5	37	3	215
LP2000329-DNA_A01	255	0	39	41	0	8	60	4	384
LP2000330-DNA_A01	1766	7	58	32	2	6	39	3	1874
LP2000331-DNA_A01	44	44	53	29	1	9	32	0	199
LP2000333-DNA_A01	347	20	42	22	1	6	33	5	448
LP6005334-DNA_B02	263	65	16	30	0	12	68	1	445
LP6005334-DNA_C02	302	1	28	25	0	7	32	2	381
LP6005334-DNA_E02	50	3	7	13	0	1	11	1	78
LP6005334-DNA_F02	466	0	149	51	0	11	65	0	700
LP6005334-DNA_F03	54	1	42	21	0	0	9	0	124
LP6005334-DNA_G01	470	56	25	29	1	13	89	1	663
LP6005334-DNA_H03	522	1	29	76	2	6	77	8	646
LP6005496-DNA_B01	67	16	14	10	0	8	51	6	162
LP6005500-DNA_A03	763	4	61	35	2	7	37	0	879
LP6005500-DNA_B01	37	1	25	31	1	3	39	1	136
LP6005500-DNA_B03	1243	88	19	15	0	8	51	3	1411
LP6005500-DNA_D02	528	2	70	47	1	14	69	5	709
LP6005500-DNA_D03	490	4	14	19	0	3	48	4	570
LP6005500-DNA_E02	54	0	17	33	0	7	69	4	171
LP6005500-DNA_F01	1353	22	44	17	1	4	40	3	1461
LP6005500-DNA_G01	425	2	30	43	1	8	55	4	534
LP6005500-DNA_G02	130	5	14	22	0	2	16	0	178
LP6005500-DNA_H01	119	5	50	30	0	15	56	3	260
LP6005500-DNA_H02	224	8	23	12	0	11	44	2	318
LP6005690-DNA_B01	691	11	17	4	0	3	40	1	756
LP6005690-DNA_B02	329	1	42	63	2	18	114	6	536
LP6005690-DNA_F03	395	4	98	11	15	10	42	2	565
LP6005690-DNA_G02	820	16	23	31	1	7	58	1	940
LP6005935-DNA_A01	287	0	37	33	0	12	96	4	444
LP6005935-DNA_A02	259	3	19	57	0	9	69	3	394
LP6005935-DNA_B03	152	39	34	28	7	10	88	7	351
LP6005935-DNA_B04	631	3	25	80	0	11	65	3	746
LP6005935-DNA_D01	13	15	54	55	1	11	70	0	206
LP6005935-DNA_D02	476	23	73	78	0	8	41	3	653
LP6005935-DNA_F01	373	28	43	52	1	3	30	1	501
LP6005935-DNA_F02	185	0	249	31	0	14	140	4	586
LP6005935-DNA_H01	117	4	31	38	0	12	76	1	253
LP6007358-DNA_A01	246	2	5	12	0	8	47	3	307
LP6007396-DNA_A01	144	2	102	43	2	11	50	2	324
LP6007407-DNA_A01	109	10	24	40	1	14	71	2	258
LP6007414-DNA_A02	29	53	17	11	0	8	28	0	144
LP6007422-DNA_A01	311	25	25	30	0	11	30	0	425
LP6007424-DNA_A01	412	1	106	97	4	13	82	2	666
LP6007427-DNA_A01	120	3	25	44	1	16	117	2	313
LP6007432-DNA_A01	309	8	23	16	3	11	64	3	428
LP6007438-DNA_A01	591	0	104	93	0	4	29	1	768
LP6007504-DNA_A01	125	0	34	34	0	3	25	1	209
LP6007518-DNA_A01	355	6	21	22	0	7	90	3	489
LP6007529-DNA_A01	80	0	18	13	1	9	59	1	171
LP6007533-DNA_A01	141	3	18	31	0	7	23	1	214
LP6007535-DNA_A01	562	2	40	42	3	4	18	1	648
LP6007540-DNA_A01	63	0	7	16	0	9	43	4	139
LP6007546-DNA_A01	35	0	9	11	0	8	32	2	89
LP6007591	73	0	33	29	1	17	81	4	224
LP6007594	587	0	52	27	1	6	67	2	717
LP6008031-DNA_C02	243	2	32	52	0	7	45	1	349
LP6008031-DNA_C03	38	0	2	0	0	0	0	0	40
LP6008031-DNA_F03	83	1	31	32	1	12	71	4	215
LP6008031-DNA_G03	144	6	68	43	4	13	104	0	363
LP6008031-DNA_H01	203	1	38	29	2	8	57	1	317
LP6008051-DNA_B02	44	0	134	34	18	6	65	1	281

LP6008051-DNA_C01	143	4	61	82	2	11	46	1	331
LP6008051-DNA_E01	68	11	15	32	0	14	55	2	185
LP6008051-DNA_F01	140	1	50	46	3	5	38	1	263
LP6008051-DNA_F02	278	18	43	20	7	20	106	3	482
LP6008141-DNA_E02	152	1	10	8	2	9	57	2	234
LP6008141-DNA_H02	169	15	60	53	8	7	63	1	328
LP6008202-DNA_A02	549	20	88	57	9	9	76	3	744
LP6008202-DNA_C01	177	6	46	75	1	12	72	11	369
LP6008202-DNA_D02	508	4	33	30	1	6	51	2	607
LP6008202-DNA_F02	0	0	13	4	1	7	41	2	64
LP6008202-DNA_H01	105	0	39	34	2	6	54	3	230
LP6008221-DNA_C01	415	0	44	40	5	6	73	2	551
LP6008221-DNA_E02	94	4	152	24	37	7	66	1	367
LP6008221-DNA_F01	256	42	23	39	3	7	42	2	399
LP6008221-DNA_G02	188	4	71	46	30	14	64	1	382
LP6008221-DNA_H01	104	35	19	22	4	10	62	5	249
LP6008280-DNA_A02	226	48	59	50	2	4	27	1	382
LP6008280-DNA_B02	158	7	16	28	1	7	25	2	231
LP6008280-DNA_C02	23	1	13	5	0	6	53	3	102
LP6008280-DNA_C03	226	3	21	51	3	9	127	2	417
LP6008280-DNA_E01	246	0	30	16	4	7	54	2	349
LP6008280-DNA_G01	47	0	4	8	0	0	4	0	60
SM-4AX84	19	0	2	5	0	9	57	2	90
SM-4B296	468	3	85	76	1	22	151	2	750
SS6003113	1	0	16	11	0	10	27	2	63
SS6003117	7	0	7	4	0	8	39	2	65
SS6003119	489	25	10	19	0	7	14	0	546
SS6003129	5	28	25	12	3	6	83	7	155
SS6003305	299	5	40	37	2	12	57	0	412
SS6003311	189	2	30	37	2	14	69	0	308
SS6003320	32	0	31	31	4	14	52	0	155
SS6003323	288	1	44	48	1	7	51	3	402

Table 3.4. Selection of best models and final list of helper genes. Shown are the parameters of the 38 unique best models in the four kernels and 24 associated unique lists of top 10 genes. These lists are named using the number of genes that compose them, followed by a letter where the same number (but not the same genes) was found multiple times. The number of times and corresponding frequency that each set of best models and list of top 10 genes was found over 500 sets and lists are also shown.

Best model (n)	Linear kernel	Radial kernel		Sigmoid kernel		Polynomial kernel			List of top 10 genes		Occurrence over 500	
	nu	nu	gamma	nu	gamma	nu	gamma	degree	n	ID	times (n)	%
1	0.05	0.05	0.03125	0.05	4	0.05	0.03125	3	1	952A	207	41.4
2	0.05	0.05	0.03125	0.05	4	0.05	0.0625	3	2	952B	161	32.2
2	0.05	0.05	0.03125	0.05	4	0.05	0.0625	3	23	952C	1	0.2
3	0.05	0.05	0.03125	0.05	4	0.05	0.125	3	2	952B	161	32.2
3	0.05	0.05	0.03125	0.05	4	0.05	0.125	3	5	951B	19	3.8
4	0.05	0.05	0.03125	0.05	4	0.05	0.25	3	2	952B	161	32.2
4	0.05	0.05	0.03125	0.05	4	0.05	0.25	3	5	951B	19	3.8
5	0.05	0.05	0.03125	0.05	4	0.05	1	3	2	952B	161	32.2
6	0.05	0.05	0.03125	0.05	4	0.05	4	3	2	952B	161	32.2
7	0.05	0.05	0.03125	0.05	4	0.05	8	3	2	952B	161	32.2
8	0.05	0.05	0.03125	0.05	4	0.05	16	3	2	952B	161	32.2
9	0.05	0.05	0.03125	0.05	4	0.05	0.015625	3	3	951A	43	8.6
9	0.05	0.05	0.03125	0.05	4	0.05	0.015625	3	10	950	3	0.6
10	0.05	0.05	0.03125	0.05	8	0.05	0.015625	3	4	934A	28	5.6
10	0.05	0.05	0.03125	0.05	8	0.05	0.015625	3	24	934B	1	0.2
11	0.05	0.05	0.03125	0.05	8	0.05	0.03125	3	4	934A	28	5.6
12	0.05	0.05	0.03125	0.05	8	0.05	0.0625	3	4	934A	28	5.6
13	0.05	0.05	0.03125	0.05	8	0.05	0.25	3	4	934A	28	5.6

14	0.05	0.05	0.03125	0.05	8	0.05	0.5	3	4	934A	28	5.6
15	0.05	0.05	0.03125	0.05	8	0.05	1	3	4	934A	28	5.6
16	0.05	0.05	0.03125	0.05	8	0.05	16	3	4	934A	28	5.6
17	0.05	0.05	0.03125	0.05	4	0.05	0.5	3	5	951B	19	3.8
18	0.05	0.05	0.03125	0.05	4	0.05	2	3	5	951B	19	3.8
19	0.05	0.05	0.03125	0.05	16	0.05	0.0625	3	6	929A	8	1.6
20	0.05	0.05	0.03125	0.05	16	0.05	0.25	3	6	929A	8	1.6
20	0.05	0.05	0.03125	0.05	16	0.05	0.25	3	25	931A	1	0.2
21	0.05	0.05	0.03125	0.05	16	0.05	2	3	6	929A	8	1.6
22	0.05	0.05	0.03125	0.05	16	0.05	4	3	6	929A	8	1.6
23	0.05	0.05	0.03125	0.05	16	0.05	16	3	6	929A	8	1.6
24	0.05	0.05	0.03125	0.05	2	0.05	0.015625	3	7	915	5	1
24	0.05	0.05	0.03125	0.05	2	0.05	0.015625	3	22	916	1	0.2
25	0.05	0.05	0.03125	0.05	2	0.05	0.25	3	7	915	5	1
26	0.05	0.05	0.015625	0.05	4	0.05	0.015625	3	8	928	4	0.8
26	0.05	0.05	0.015625	0.05	4	0.05	0.015625	3	9	929B	3	0.6
26	0.05	0.05	0.015625	0.05	4	0.05	0.015625	3	18	931B	1	0.2
27	0.05	0.05	0.015625	0.05	4	0.05	8	3	8	928	4	0.8
28	0.05	0.1	0.015625	0.05	16	0.05	0.25	3	11	920	3	0.6
29	0.05	0.05	0.015625	0.05	16	0.05	0.015625	3	12	926A	2	0.4
30	0.05	0.05	0.0078125	0.05	2	0.05	0.25	3	13	898	1	0.2
31	0.05	0.05	0.0078125	0.05	16	0.05	0.25	3	14	911	1	0.2
32	0.05	0.05	0.015625	0.05	0.5	0.05	0.0625	3	15	907	1	0.2
33	0.05	0.05	0.015625	0.05	1	0.05	0.0625	3	16	909	1	0.2
34	0.05	0.05	0.015625	0.05	2	0.05	0.25	3	17	906A	1	0.2

35	0.05	0.05	0.015625	0.05	8	0.05	0.015625	3	19	926B	1	0.2
36	0.05	0.05	0.03125	0.05	0.5	0.05	0.015625	3	20	906B	1	0.2
37	0.05	0.05	0.03125	0.05	1	0.05	0.25	3	21	919	1	0.2
38	0.05	0.1	0.015625	0.05	2	0.05	0.015625	3	26	908	1	0.2

3.3.3 Feature correlation

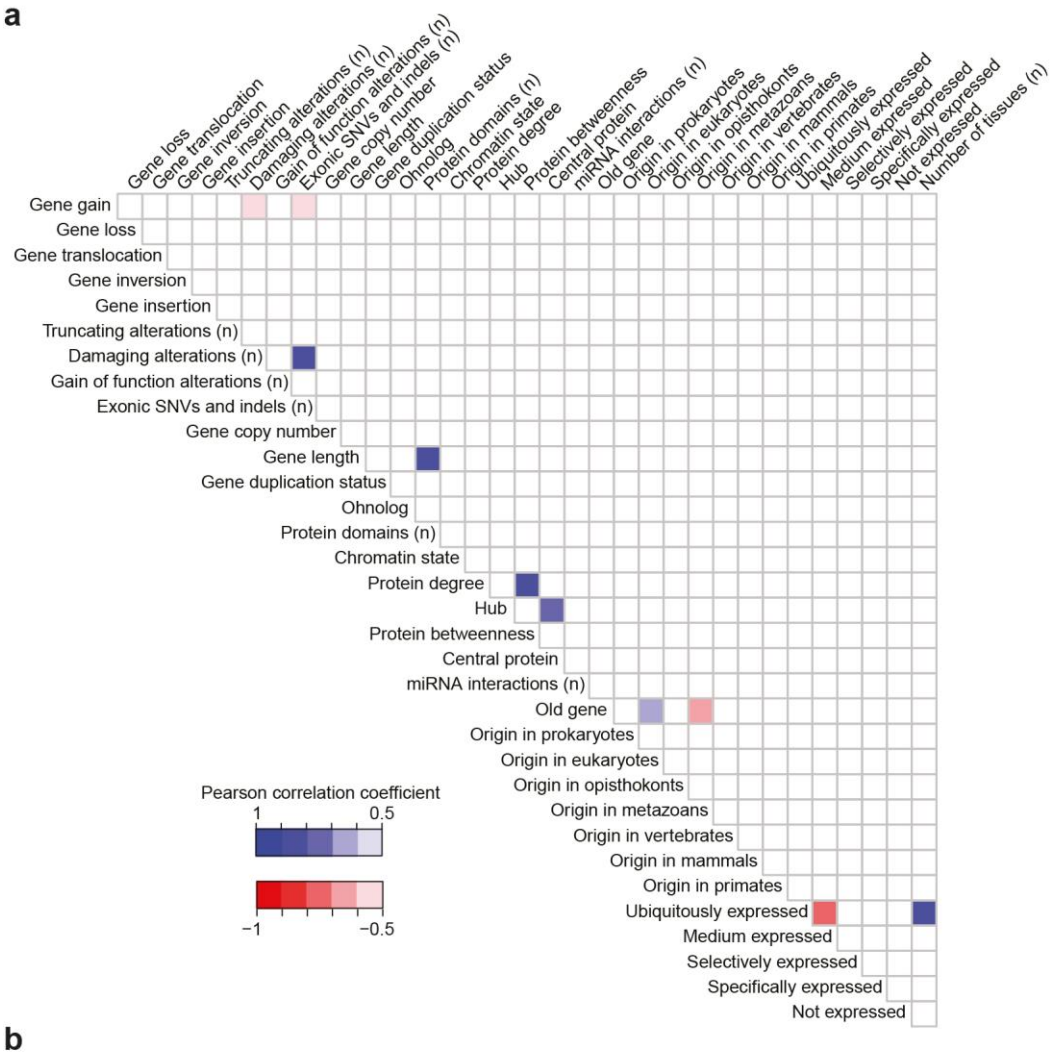
All sysSVM features exhibit statistically significant difference between known cancer genes (as those are defined in the Cancer Gene Census) and the rest of human genes (see chapter 2). However, the correlation of these features was not thoroughly explored during the pilot phase of my thesis. To this end, I performed a pairwise correlation analysis of the sysSVM features (Figure 3.5A).

Overall, I observed positive correlations between i) the number of exonic mutations (mutational burden) and the number of damaging mutations, ii) the protein degree and betweenness, and iii) the gene length and the number of protein domains. Moreover, several sysSVM features were mutually exclusive, and therefore exhibited high negative correlation (Figure 3.5A). These features included various expression measurements derived from GTEx (https://www.gtexportal.org/home/) across human tissues, and features related to the evolutionary origin of human genes. As it is easily conceivable, a gene cannot be ubiquitously expressed and not expressed at the same time. Cancer genes have been found to be expressed in a significantly higher number of normal tissues than the rest of human genes, thus, features representing low or no expression in normal tissues were mostly constant in the OAC training set.

The high positive correlation of the damaging mutations with the mutation burden (SNVs and indels) suggested that these features carry some degree of redundancy. However, I argued that although highly correlated, the two features would be relevant to sysSVM, if their ratio exhibited significant difference between known cancer genes and the rest of human genes. To avoid biased results on the OAC training set, this analysis was performed pan-cancer. To this end, I downloaded level 3 whole-exome mutation data from The Cancer

Genome Atlas (TCGA; data downloaded on 01/03/2015) for 7,828 tumour samples covering 31 cancer types (Table 3.5). TCGA somatic mutations were re-annotated using an in-house mutation annotation pipeline in order for variant calls across different cancer types to be comparable (see Methods). The total number of mutations analysed was 579,699 and 1,237,039 for silent and non-silent mutations, respectively (Table 3.5). I found that known cancer genes had significantly higher ratio of damaging mutations to all mutations in the pan-cancer analysis (Figure 3.5B), and therefore I decided to retain the number of all exonic mutations in the feature list of sysSVM. Although such “ratiometric” gene features have been used before to identify driver genes (Tokheim et al. 2016), I acknowledged that their contribution to sysSVM was expected to be minimal. This is because, in sysSVM, each gene in each patient is considered separately and therefore the number of mutations for both exonic and damaging mutations per gene will be low (mostly 1).

Protein degree and betweenness were also very strongly correlated. It has been previously shown that network centrality measurements, such that of degree and betweenness, are correlated in several complex networks (Lee 2006). However, they capture different characteristics of the network. Degree describes the number of primary neighbours of a vertex, while betweenness is a measure of the influence of a vertex over the flow of information between every pair of vertices, under the assumption that information primarily flows over the shortest path between them. Therefore, both were retained in the feature list of sysSVM for this part of my thesis. Finally, as gene length and the number of protein domains exhibited significant difference between known cancer genes and the rest of human genes, both features were maintained in the feature list of sysSVM for this analysis.


a
Gene loss
Gene translocation
Gene inversion
Gene insertion
Truncating alterations (n)
Damaging alterations (n)
Gain of function alterations (n)
Exonic SNVs and indels (n)
Gene copy number
Gene length
Gene duplication status
Ohnolog
Protein domains (n)
Chromatin state
Protein degree
Hub
Protein betweenness
Central protein
miRNA interactions (n)
Old gene
Origin in prokaryotes
Origin in eukaryotes
Origin in opisthokonts
Origin in metazoans
Origin in vertebrates
Origin in mammals
Origin in primates
Ubiquitously expressed
Medium expressed
Selectively expressed
Specifically expressed
Not expressed
Number of tissues (n)
Pearson correlation coefficient
1
0.5
−1
−0.5

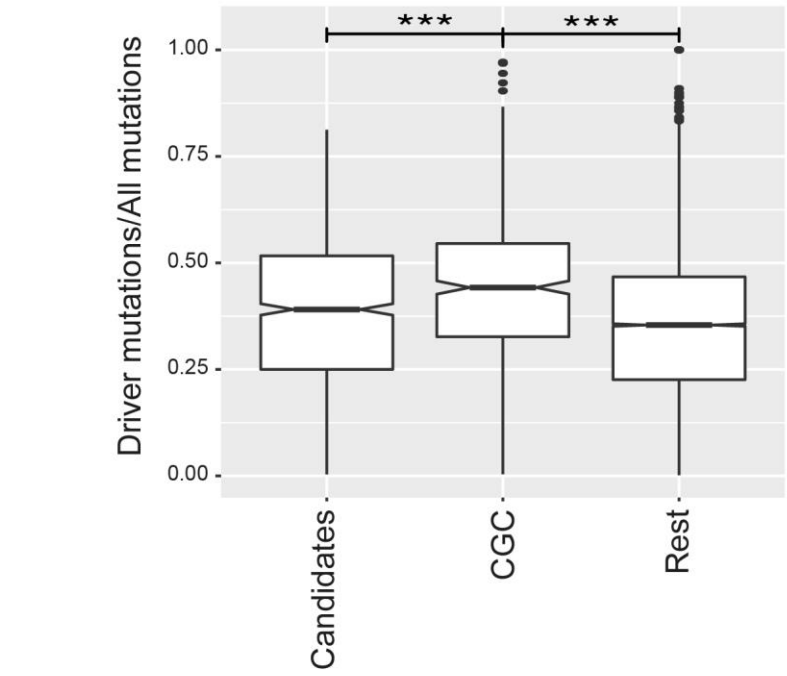

b
***
***
Driver mutations/All mutations
1.00
0.75
0.50
0.25
0.00
Candidates
CGC
Rest

Figure 3.5. Correlation of sysSVM features. **(a)** Pairwise correlation matrix of sysSVM features. Blue denotes positive pearson correlation coefficient while red denotes negative correlation coefficient. Only coefficients higher than 0.5 or lower than -0.5 are shown. White colour denotes coefficients between -0.5 and 0.5. **(b)** Distribution of the ratio of predicted damaging mutations over all exonic mutations for known cancer genes, candidate cancer genes and the rest of human genes. Known cancer genes were derived from the Cancer Gene Census (Tate et al. 2018). Candidate cancer genes were derived from the Network of Cancer Genes database (An, Dall'Olio, et al. 2016) and they represent genes that have been predicted as drivers by various methods, but their driver role is pending further validation. Finally, the rest of human gene set is comprised of all human genes that were neither in the known cancer genes nor in the candidate cancer genes. (*** $p < 0.01$).

Table 3.5. Pan-cancer cohorts from The Cancer Genome Atlas. Shown are the number of patients, and the number of silent and non-silent mutations for 31 cancer types.

Cancer type	Abbreviation	Patients	Silent mutations	Non-silent mutations
Adrenocortical Carcinoma	ACC	72	2,538	3,905
Bladder Urothelial Carcinoma	BLCA	232	19,322	46,723
Breast Invasive Carcinoma	BRCA	954	20,108	42,780
Cervical squamous cell carcinoma and endocervical adenocarcinoma	CESC	179	26,648	41,343
Cholangiocarcinoma	CHOL	35	1,968	3,227
Colon adenocarcinoma	COAD	255	20,956	52,773
Oesophageal carcinoma	ESCA	179	27,687	45,685
Glioblastoma multiforme	GBM	143	2,898	7,576
Head and Neck squamous cell carcinoma	HNSC	491	37,931	86,993
Kidney Chromophobe	KICH	65	2,401	5,097
Kidney renal clear cell carcinoma	KIRC	423	8,575	21,241
Kidney renal papillary cell carcinoma	KIRP	164	7,356	18,099
Acute Myeloid Leukaemia	LAML	169	616	1,534
Brain Lower Grade Glioma	LGG	506	7,936	17,547
Liver hepatocellular carcinoma	LIHC	187	35,222	55,802
Lung adenocarcinoma	LUAD	487	77,086	211,964
Lung squamous cell carcinoma	LUSC	174	14,596	42,471
Ovarian serous cystadenocarcinoma	OV	341	5,283	13,952
Pancreatic adenocarcinoma	PAAD	136	8,758	11,941
Pheochromocytoma and Paraganglioma	PCPG	175	2,359	4,824

Prostate adenocarcinoma	PRAD	409	4,267	8,634
Rectum adenocarcinoma	READ	110	3,010	7,575
Sarcoma	SARC	237	16,650	25,032
Skin cutaneous melanoma	SKCM	357	142,039	249,273
Stomach adenocarcinoma	STAD	335	47,526	128,524
Testicular Germ Cell Tumours	TGCT	143	2,977	4,537
Thyroid carcinoma	THCA	393	1,154	3,116
Thymoma	THYM	116	5,863	10,623
Uterine Corpus Endometrial Carcinoma	UCEC	229	22,832	59,180
Uterine Carcinosarcoma	UCS	53	1,631	2,587
Uveal melanoma	UVM	79	1,506	2,481
Total		**7,828**	**579,699**	**1,237,039**

3.3.4 Distribution of known cancer genes in the feature space

After I established that sysSVM features were descriptive of known cancer genes (as they exhibited statistically significant difference between known cancer genes and the rest of human genes), I sought to investigate whether specific combinations of these features can define classes of training observations in the feature space of sysSVM. To this end, I performed a dimensionality reduction analysis using t-distributed stochastic neighbour embedding (t-SNE) and measured the average distance of known cancer genes from their closest high-density peak in the t-SNE map. I found that there were multiple clusters of known cancer genes, each defining a certain region of the feature space (Figure 3.6A). Moreover, permutation analysis showed that these clusters were not random (Figure 3.6B).

When specific features were mapped onto the 2-dimensional space, it became evident that each high-density region (as those defined by the training observations in Figure 3.6A) represented genes harbouring a particular sysSVM feature (Figure 3.7). For instance, genes with high degree in the protein-protein interaction network, translocations, inversions, and truncating mutations were clustered together; in contrast several other features, such as amplifications, the breadth of expression, and the chromatin compartment were uniformly distributed across the whole feature space (Figure 3.7). Of note, such features with uniform distribution would be non-informative for a two-class classifier. This is because in a two-class setting only the features that maximise the differences between the two classes are of interest. Conversely, uniformly distributed features are informative in one-class classification algorithms, such as sysSVM, owing to the presence of only one class of training observations (positive observations).

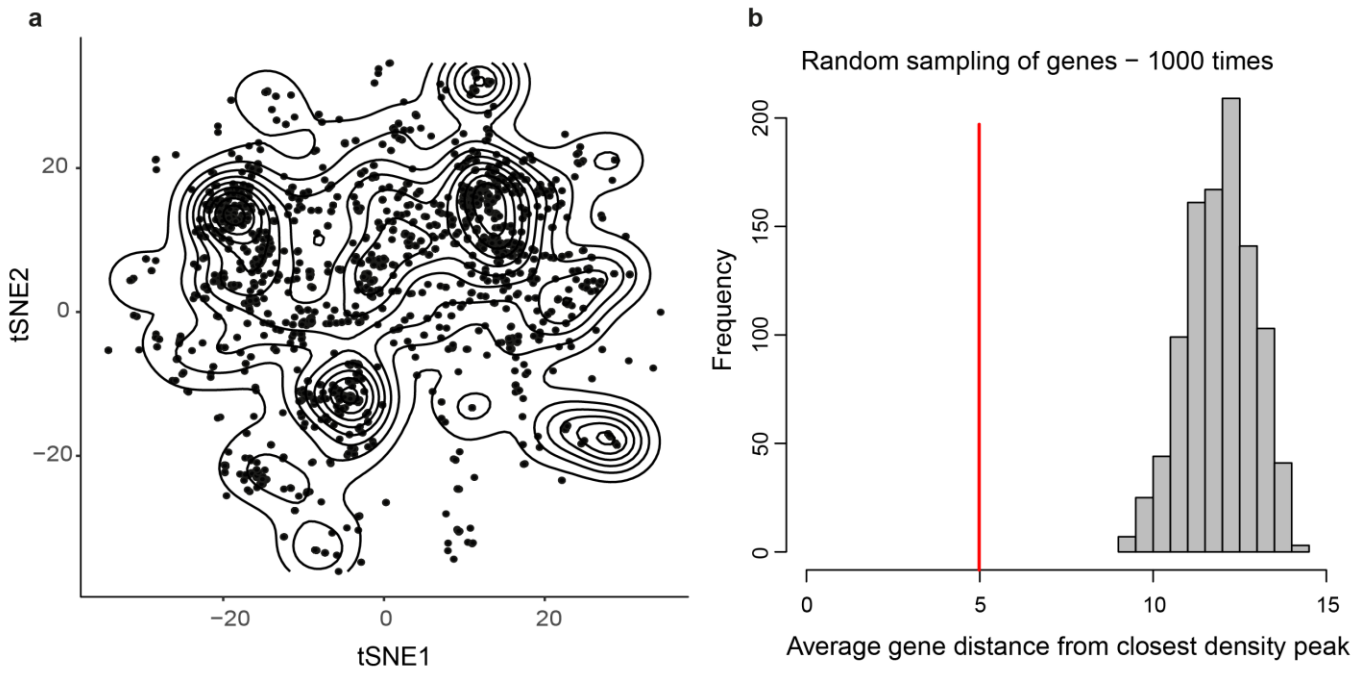


Figure 3.6. Clustering of the training observations in the feature space of sysSVM. **(a)** Dimensionality reduction of the training set of sysSVM using t-distributed stochastic neighbour embedding (tSNE). **(b)** Permutation analysis of the clustering of the training observations in a. To assess whether the training observations were clustered more than expected in the high-density regions of the 2-D feature space, I randomly sampled 4,091 genes (equal size as the training set) 1,000 times. After defining the peaks of their distribution in the 2-D feature space, I measured the distance of each gene from the closest density peak and averaged the distance of all genes within each iteration. The mean of gene distance for the random samples was significantly higher than the true average distance of the training observations from their density peaks (red vertical line).

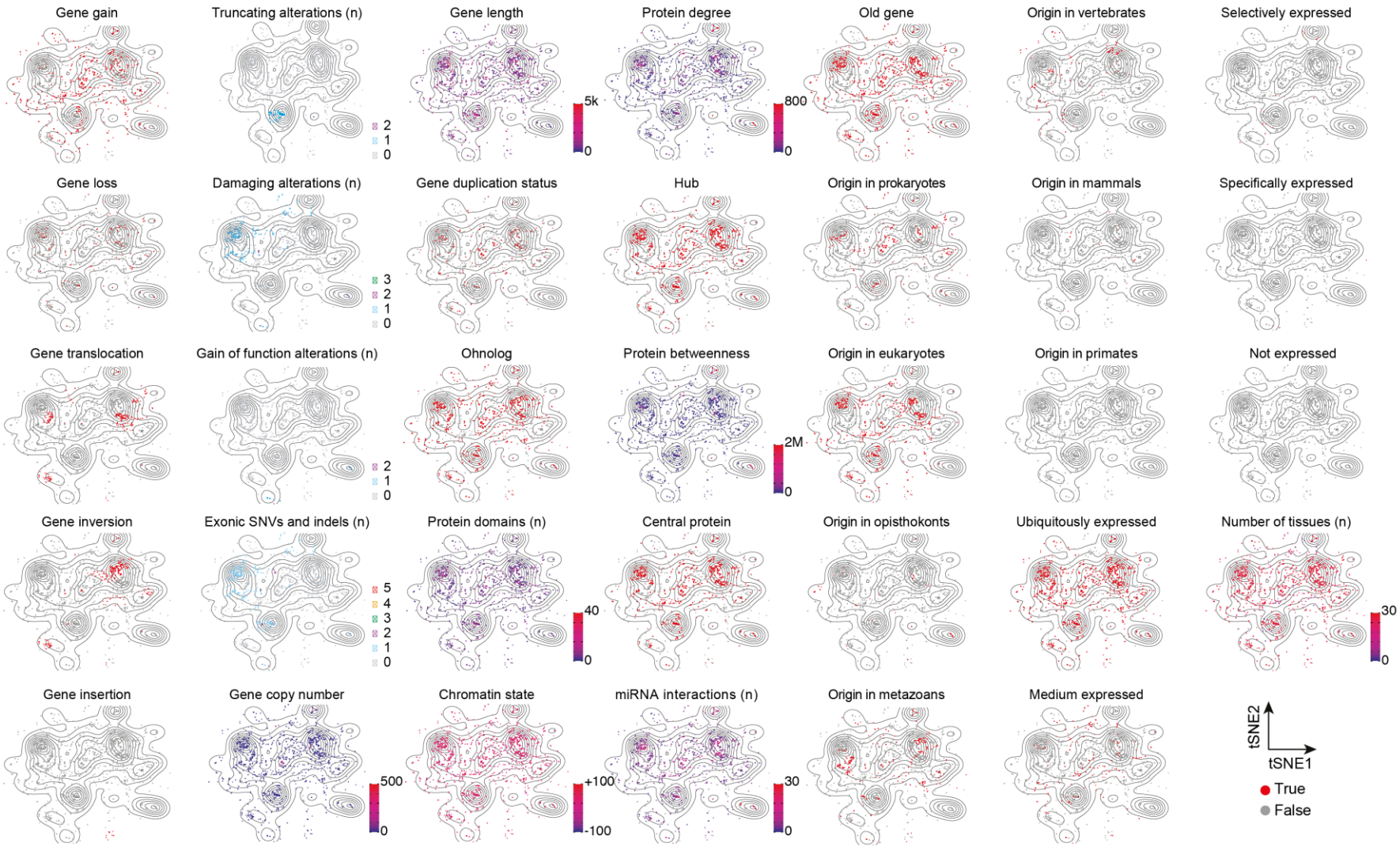

Gene gain
Truncating alterations (n)
2
1
0
Gene length
5k
0
Protein degree
Old gene
800
0
Origin in vertebrates
Selectively expressed
Gene loss
Damaging alterations (n)
3
2
1
0
Gene duplication status
Hub
Origin in prokaryotes
Origin in mammals
Specifically expressed
Gene translocation
Gain of function alterations (n)
2
1
0
Ohnolog
Protein betweenness
2M
0
Origin in eukaryotes
Origin in primates
Not expressed
Gene inversion
Exonic SNVs and indels (n)
5
4
3
2
1
0
Protein domains (n)
40
0
Central protein
Origin in opisthokonts
Ubiquitously expressed
Number of tissues (n)
30
0
Gene insertion
Gene copy number
500
0
Chromatin state
miRNA interactions (n)
+100
-100
Origin in metazoans
30
0
Medium expressed
tSNE2
tSNE1
True
False

Figure 3.7. For each property, a 2-D map of the high-dimensional data was rebuilt for the 476 known cancer genes altered 4,091 times in the cohort of 261 OACs. Black curves represent the density of known cancer genes. For continuous or multi-value variables, a colour code is reported. For categorical variables genes are labelled according to whether they have (red) or not (grey) that property. All properties are summarised in Table 3.2.

3.3.5 Description of the best models

Confirming the observations in the pilot cohort (Table 2.4), the best value of the parameter *nu* for all kernels in the extended OAC cohort was 0.05, suggesting that at least 95% of the training set was considered for the construction of the decision boundary during the training phase. Gamma, although it fluctuated between three possible values (Table 3.4), converged to the value of 0.03125, with the exception of the sigmoid kernel in which the best value was 4. Finally, the best parameter for degree in the polynomial kernel was 3.

Next, I examined the distribution of the decision values in the best models. As noted in chapter 2, decision values correspond to the distance from the decision boundary and in sysSVM they are used to calculate the meta-score used to rank the genes (equation 2.10). Overall, very few genes in the training set had negative decision values (Figure 3.8), confirming the high sensitivity that was estimated during the cross-validation. Specifically, linear and sigmoid kernels had the highest average decision values (75.35 and 121.85, respectively; Figure 3.8) and the lowest number of support vectors (214 and 205, respectively).

To examine the weights of individual features in the final models, I performed a recursive feature elimination analysis (Guyon et al. 2002). Briefly, the weight vector (w) of the features in each model was computed by multiplying the sysSVM

coefficients with the support vectors and all features were ranked according to the second power of w (for details see chapter 2). Recursive feature elimination was run 33 times per kernel to estimate the ranks of all 34 features. In contrast to the pilot cohort, copy number gains was not the feature with the highest weight in any of the kernels (Table 3.6). Instead, it was ranked fifth and sixth in linear and polynomial kernels, respectively. Other top-ranking features were the expression of genes in human tissues, and the number of connections in the protein-protein interaction network (Table 3.6). Interestingly, high-ranking features were different across different kernels, in contrast to the more homogeneous ranking in the pilot cohort. This suggested that the low number of training observations in the pilot phase affected the ranking of the features in the best models. Interestingly, radial and sigmoid kernels assigned higher weights to degree, betweenness and gain-of-function mutations, while linear and polynomial kernels weighted more expression, origin and the age of the genes (Table 3.6). Taken together, these results highlight that molecular features were ranked on average higher in the best model of the extended cohort than in the pilot cohort, despite the fact that the weight of copy number gains was decreased.

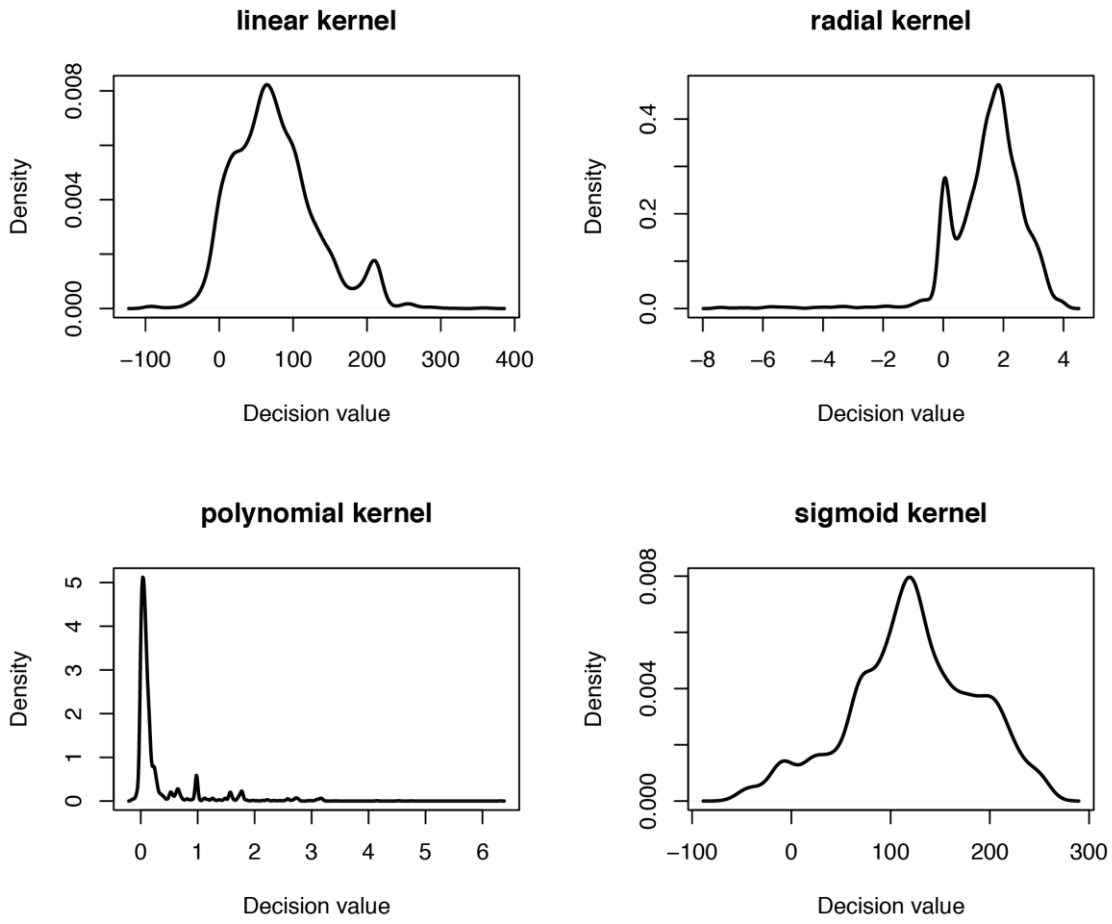


Figure 3.8. Distribution of decision values of the training observations in the best models.

Table 3.6. Ranks of systems-level (blue) and molecular (orange) features in sysSVM as derived from recursive feature elimination for each kernel.

Rank	Linear	Polynomial	Radial	Sigmoid
1	Medium expressed	Ubiquitously expressed	Protein degree	Protein betweenness
2	Ubiquitously expressed	Medium expressed	Protein betweenness	Protein degree
3	Origin in vertebrates	Origin in vertebrates	Gain-of-function mutations	miRNA interactions
4	Origin in metazoans	Origin in metazoans	miRNA interactions	Gain-of-function mutations
5	Gene gain	Old gene	Copy number	Ubiquitously expressed
6	Old gene	Gene gain	Hub	Ohnolog
7	Selectively expressed	Selectively expressed	Central protein	Central protein
8	Ohnolog	Ohnolog	Ubiquitously expressed	Hub
9	Central protein	Central protein	Ohnolog	Origin in opisthokonts
10	Origin in prokaryotes	Origin in eukaryotes	Old gene	Old gene
11	Origin in eukaryotes	Origin in prokaryotes	Truncating mutations	Damaging mutations
12	Origin in mammals	Origin in mammals	Gene gain	Exonic SNVs
13	Gene loss	Gene loss	Origin in metazoans	Origin in eukaryotes
14	Origin in opisthokonts	Origin in opisthokonts	Exonic SNVs	Origin in metazoans
15	Specifically expressed	Specifically expressed	Origin in eukaryotes	Origin in prokaryotes
16	Hub	Hub	Gene loss	Truncating mutations
17	Not expressed	Not expressed	Origin in opisthokonts	Medium expressed
18	Origin in primates	Origin in primates	Gene duplication	Gene loss
19	Gene duplication	Gene duplication	Origin in prokaryotes	Gene duplication
20	Exonic SNVs	Copy number	Medium expressed	Origin in vertebrates
21	Gene translocation	Protein degree	Origin in vertebrates	Gene gain
22	Gene inversion	Damaging mutations	Selectively expressed	Specifically expressed
23	Protein degree	Chromatin state	Specifically expressed	Selectively expressed
24	Damaging mutations	Gene length	Origin in mammals	Not expressed
25	Truncating mutations	Gene translocation	Not expressed	Origin in primates
26	Chromatin state	Gene inversion	Origin in primates	Origin in mammals
27	miRNA interactions	Protein betweenness	Damaging mutations	Gene length
28	Protein domains	Protein domains	Gene insertion	Gene translocation
29	Gain-of-function mutations	Gene insertion	Gene inversion	Gene inversion
30	Gene length	Gain-of-function mutations	Number of tissues	Chromatin state
31	Gene insertion	miRNA interactions	Gene length	Copy number
32	Protein betweenness	Truncating mutations	Gene translocation	Number of tissues
33	Copy number	Exonic SNVs	Chromatin state	Gene insertion
34	Number of tissues	Number of tissues	Protein domains	Protein domains

3.3.6 The landscape of patient-specific cancer genes in OAC

My working hypothesis was that the driver potential of genes with predicted damaging alterations could be predicted by the similarity of their properties to those of known cancer genes and their relative contribution to tumorigenesis declined with decreasing sysSVM score. Therefore, I considered the top scoring genes in each patient as the most likely contributors to cancer progression. Overall, these genes localised closely to the high-density regions of known cancer genes (Figure 3.9A). Furthermore, using the mean distance from the centre of each high-density area in the tSNE map, I verified that the top scoring genes occupied positions proximal to those of known cancer genes, (Figure 3.9B). This indicated that the properties of top scoring genes indeed resembled those of known cancer genes, and that the scoring function accurately recapitulated the distance from the decision boundary in each kernel.

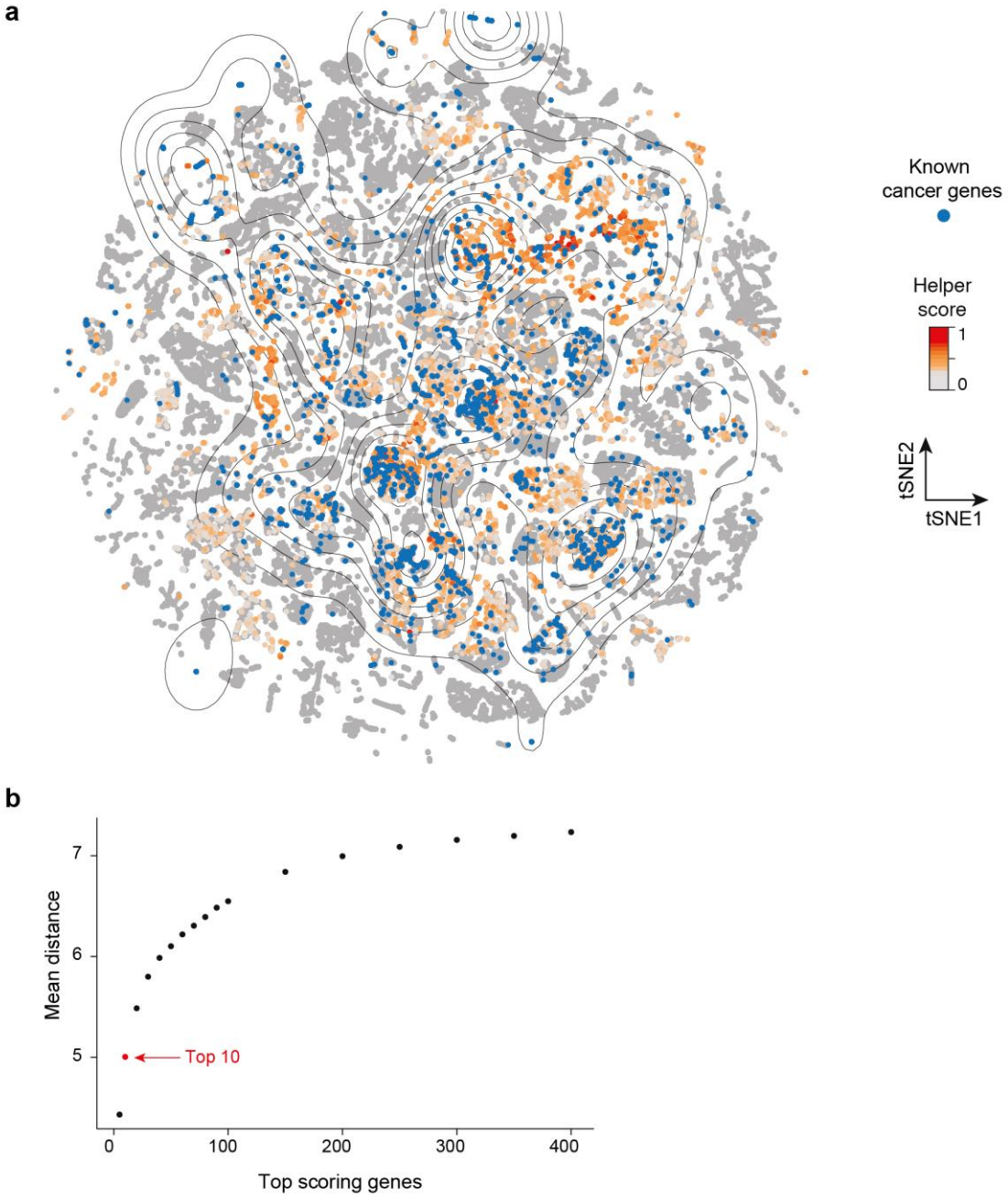


Figure 3.9. Overview of altered genes in the 261 OACs in the feature space. (**a)** t-distributed Stochastic Neighbour Embedding (t-SNE) plot of 116,989 altered genes in 261 OACs. Starting from the 34 properties used in sysSVM, a 2-D map of the high-dimensional data was built using Rtsne package in R. Curves are coloured according to the density of 4,091 known cancer genes (blue dots) used as a training set and the rest of altered genes are coloured according to their sysSVM score. **(b)** Average distance from the center of the highest-density regions of known cancer genes for different score thresholds to define helper genes. For each

gene set, the distance of each helper gene from the center was calculated and the mean of the distribution across the set was derived.

To evaluate how the sysSVM classifier, trained on the ICGC cohort, performed on independent cohorts, I used 86 OACs from The Cancer Genome Atlas (TCGA) and 21 OACs from a previous study (Nones et al. 2014) (Table 3.7). I scored all altered genes, including known cancer genes, in each of the 107 OACs independently, using the four classifiers trained on the ICGC cohort. In both datasets, known cancer genes (*e.g. TP53*, *ERBB2, EGFR*, *SMAD4*) had significantly higher scores than the rest of the altered genes (Figure 3.10), indicating that sysSVM was able to recognise them as major cancer contributors in previously unseen cancer samples.

To further investigate the top sysSVM predictions in the 261 OACs, I analysed the 952 helper genes (top 10 scoring genes in each patient; Appendix Table 7.2). As sysSVM uses no pre-defined cut-off to distinguish driver from passenger genes, I also checked whether the main findings hold true when a higher or lower number of top scoring genes were considered (see chapter 5 for the analysis of the alternative cut-offs). The vast majority (nearly 80%) of newly predicted helpers underwent copy number gain (Figure 3.11A), consistently with the prevalence of gene amplification in OAC (Table 3.2). To examine the consequence of helper amplification, I analysed the expression of amplified helpers in 92 OACs for which expression data were available. Amplified helpers were found over-expressed in OACs in which they were amplified, as compared to OACs in which they had a copy neutral state (Figure 3.11B). This suggested that

amplification of these helpers may have functional implications in OAC. Moreover, approximately 60% of helper genes were rare or patient-specific (Figure 3.11C). A few helper genes, however, were altered in more than 5% of OACs (Table 3.8) and their relatively higher frequency of alteration could not be explained by large-scale amplification events containing neighbouring known drivers (i.e. concomitant amplification of the closest driver gene; Figure 3.11D) (Secrier et al. 2016). Overall, 171 helpers (18%) had been predicted as candidate cancer genes in previous studies (Repana et al. 2018), and 41 helpers (4%) were recently added to Tiers 1 and 2 of the Cancer Gene Census (Tate et al. 2018), indicating that their tumorigenic potential has been validated[3].

The most recurrently altered helper was *TOMM34*, a translocase of the outer membrane of mitochondria, which was found amplified in 33 OACs (12.6%; Table 3.8). *TOMM34* is an essential factor for protein import in mitochondria, and it has been shown to interact with the mature portion of several preproteins during their translocation through the mitochondrial membrane (Nuttall et al. 1997). Overexpression of *TOMM34* has been reported in colon (Shimokawa et al. 2006) and early invasive breast cancer (Aleskandarany et al. 2012), suggesting that mitochondrial dysfunction plays a role in tumorigenesis. In particular, protein expression of *TOMM34* has been associated with higher tumour grade, advanced nodal stage and larger tumour size in breast cancer, demonstrating the utility of this gene as a potential biomarker. Although, no difference in tumour grade or nodal status was observed in OAC, *TOMM34* was predicted as helper in almost all the samples in which it was amplified (33/35). Other recurrent helpers (*NCOA3*,

[3] Intersection of helpers with genes from previous studies was performed by Damjan Temelkovski and Joel Nulsen

E2F1, *MCM7*, *VAPB*, *DNMT3B*) have been reported to play a role in tumour progression in multiple tissues (Wagner et al. 2013; Liang et al. 2016; Qu et al. 2017; Rao et al. 2012; Peralta-Arrieta et al. 2017). A more comprehensive view of all predicted helpers, along with data on their experimental validation is presented in chapter 4.

To analyse the impact of false positive predictions in sysSVM, I collected two lists of previously described false positive drivers from the literature. The first list was composed of 49 genes that had been previously considered as false positive cancer driver genes (Lawrence et al. 2013; An et al. 2016). Overall, only three of these genes (*PCLO*, *CNTNAP2* and *NRXN3*) were predicted by sysSVM to be cancer helpers. *PCLO* has been considered as a false positive, because of its long coding sequence and biased base composition (Lawrence et al. 2013). However, *PCLO* has been shown to exert an oncogenic role in oesophageal cancer by interfering with the *EGFR* signalling pathway (Zhang et al. 2017). The second list was a manually curated set of 488 genes (Bailey et al. 2018). I found 44 helpers in this list (4.6% of the total). This fraction was smaller than the fraction of known cancer genes (Futreal et al. 2004) that were present in the same list of false positives (46/719, 6.4%). Altogether, these analyses indicate that sysSVM robustly predicts cancer genes in multiple patient cohorts, with minimal false positive rate.

Patient-specific analysis of helpers was also conducted to investigate whether the 2-dimensional feature space (Figure 3.6A) could be utilised to infer patient sub-groups, based on which high-density area the corresponding helpers were localised in. As the location of helpers in the 2-D feature space reflects their similarity to known cancer genes, this analysis could highlight groups of patients with strong, intermediate and weak cancer helpers. Overall, three types of OACs

were identified. Those that had all their helpers clustered in one region (Figure 3.12A), those with the majority of the helpers positioned in multiple high-density regions (Figure 3.12B) and, finally, those with helpers located in sparse areas of the feature space (Figure 3.12C). I reasoned that an overall high similarity of helpers to the major drivers (Figure 3.12A and 3.12B) could serve as a biomarker for tumour aggressiveness. Conversely, a dispersed pattern of helpers, away from the high-density areas (Figure 3.12C), would indicate helpers with lower driver potential and, therefore, possibly a less aggressive tumour. However, I found no evidence of association of sysSVM score with survival when examined OACs with high-scored helpers versus OACs with low-scored helpers (Figure 3.12D) or when I examined sysSVM score regardless of the cut-off of top-scoring genes (Figure 3.12E).

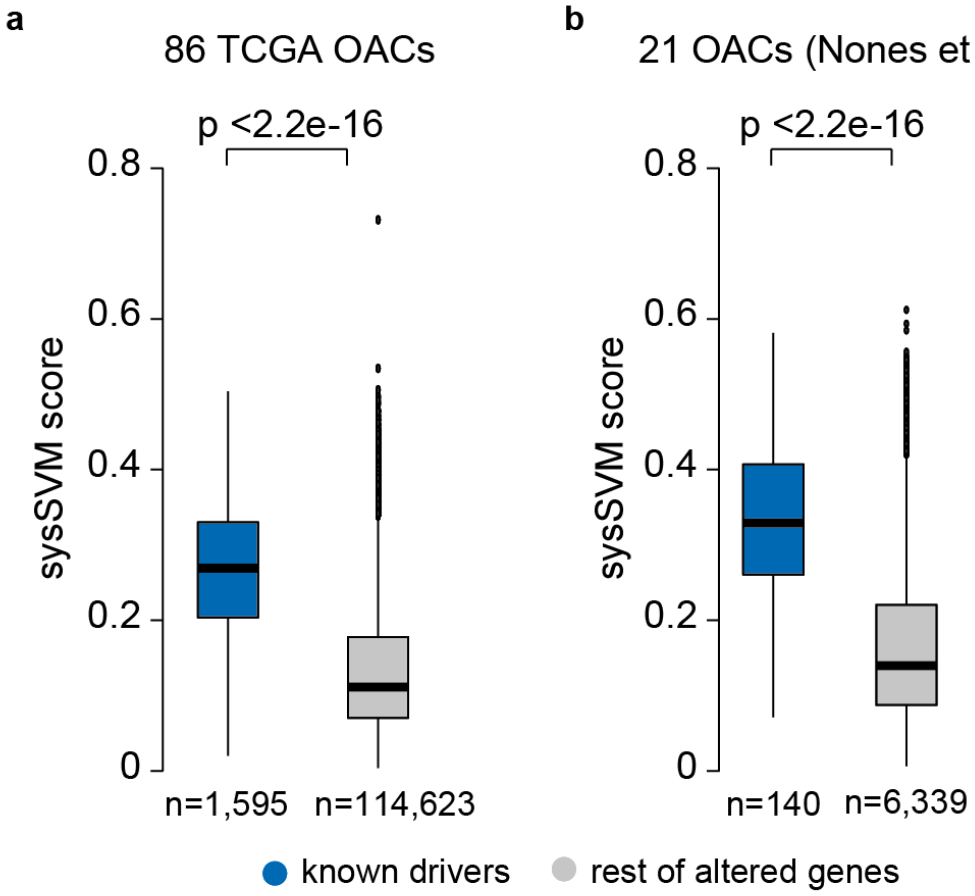


Figure 3.10. Comparison of sysSVM scores between known drivers and the rest of altered genes for 86 OACs from TCGA (**a**) and 21 OACs from Nones et al. (Nones et al. 2014) (**b**). Starting from all altered genes, known drivers were identified as described in the Methods. All genes that are not expressed in healthy esophagus were removed from both gene sets. Distributions were compared using two tailed Wilcoxon rank-sum test.

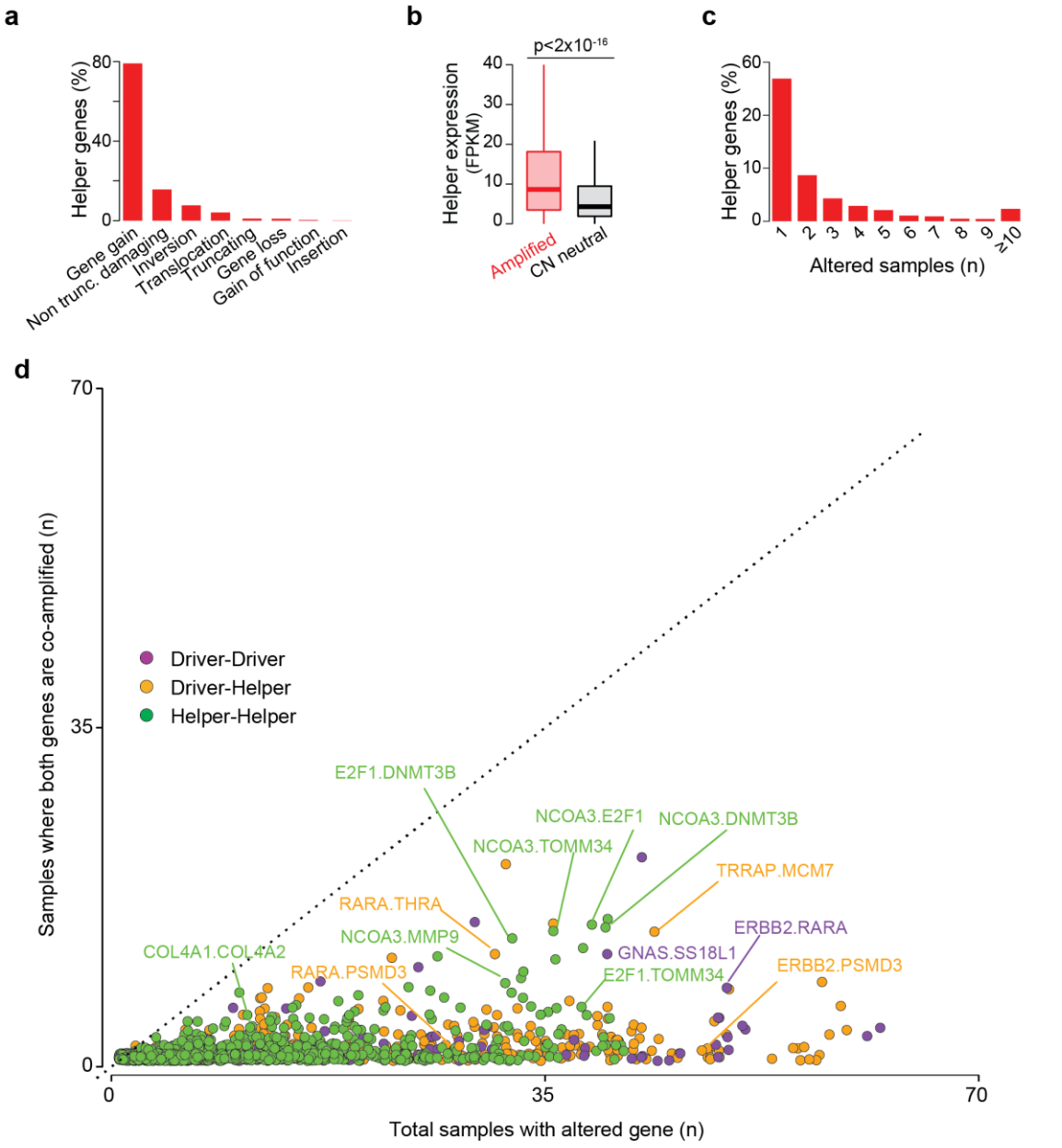


Figure 3.11. Characteristics of cancer helpers. **(a)** Distribution of damaging alterations in 952 cancer helpers. Overall, these genes acquire 2,608 damaging alterations. **(b)** Comparison of the expression of cancer helpers in OACs in which they have been found amplified with OACs in which they were not predicted as helpers and they were copy number neutral. **(c)** Recurrence of cancer helpers across 261 OACs. Only samples acquiring alterations with a damaging effect are considered (see main text). **(d)** Scatterplot of co-amplified driver and helper genes in 261 OACs as a function of the total number of samples where they are altered. For each pair of drivers or helpers in the same chromosome, ASCAT breakpoints were used to assign whether they were in the same co-amplified segment.

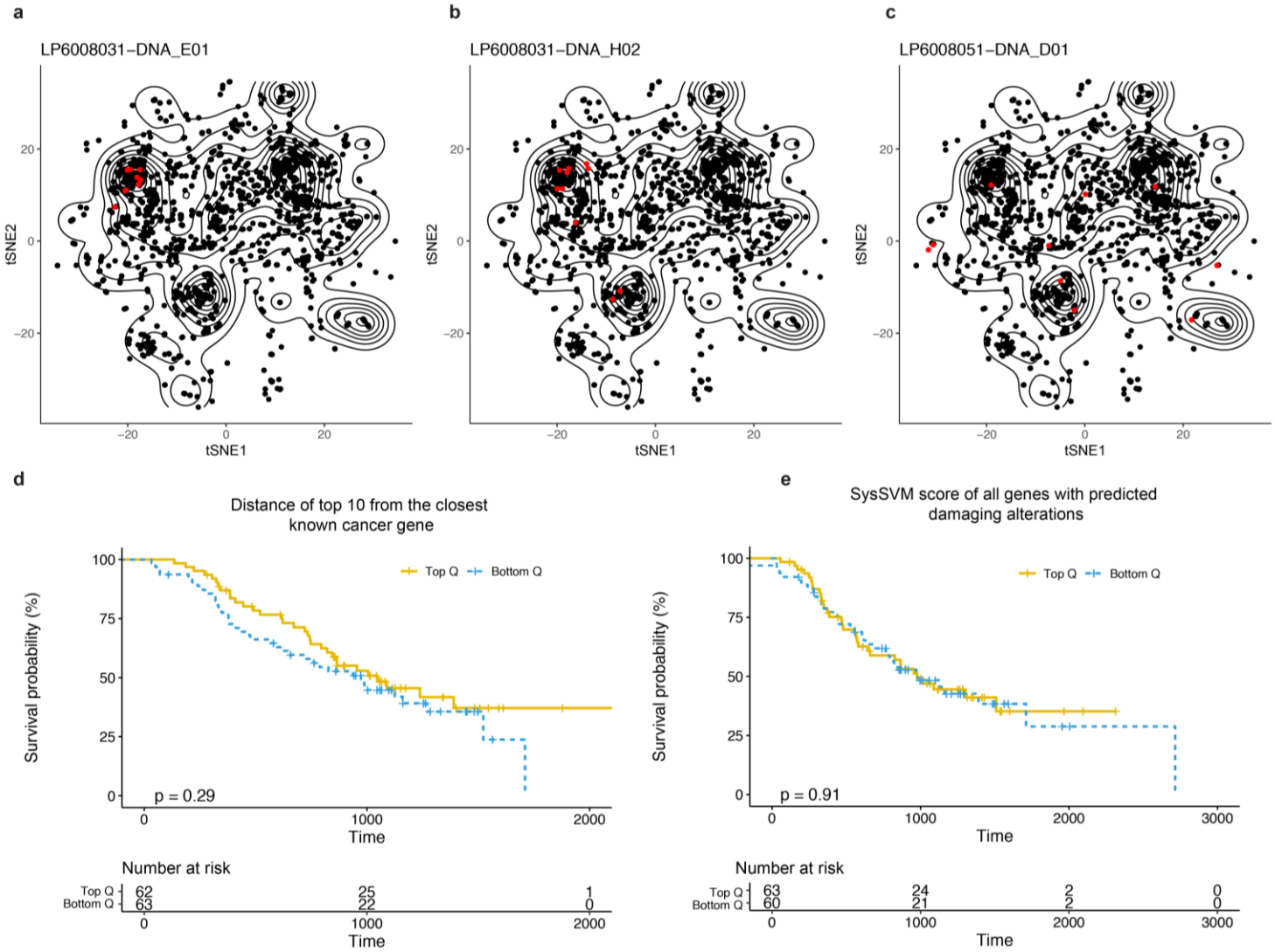

a
LP6008031−DNA_E01
b
LP6008031−DNA_H02
c
LP6008051−DNA_D01
tSNE1
tSNE2
d
Distance of top 10 from the closest known cancer gene
Survival probability (%)
Top Q
Bottom Q
p = 0.29
Time
Number at risk
Top Q 62 25 1
Bottom Q 63 22 0
e
SysSVM score of all genes with predicted damaging alterations
p = 0.91
Top Q 63 24 2 0
Bottom Q 60 21 2 0

Figure 3.12. Similarity of cancer helpers with known driver genes within OACs. Representative examples of OACs: **(a)** all helpers concentrated in one region of the 2-D feature space; **(b)** all helpers concentrated in multiple high-density regions; **(c)** a more dispersed distribution of helpers. **(d)** Survival analysis of OACs with helpers located very close to known drivers versus OACs with helpers located further away from known drivers in the 2-D feature space. For each helper in each OAC, the distance of the closest known driver was calculated, and an average distance of all helpers was computed per OAC. OACs in the upper and bottom quartiles of the average distance were compared. **(e)** Same as d but for all scored genes within each OAC.

Table 3.7. Summary of somatically altered genes in the two validation cohorts (86 OACs from The Cancer Genome Atlas and 21 OACs from previous literature). For each of the 86 TCGA and 21 OACs, the number of genes with predicted damaging alterations is provided. The total number of altered genes is 155,886 (17,779 unique hits), and 8,800 (5,617 unique hits) in the two datasets, respectively.

Sample	Set	Genes with damaging alterations (n)								
		Gain	Loss	Trans location	Inversion	Insertion	Truncating	Non-truncating damaging	Gain of function	Total
03aa1df3-5156-4bfd-a36d-4daff2b6d06c	Nones et al 2014	227	23	0	0	0	19	50	0	319
09681159-48b2-4607-b455-4284c49b49a5	Nones et al 2014	593	10	0	0	0	16	62	0	678
2206848d-3794-4061-a379-3eb1582e4ea5	Nones et al 2014	815	9	0	0	0	6	26	0	853
430b3d28-2f55-4cd8-91ec-3b94cadb34b8	Nones et al 2014	531	0	0	0	0	8	72	0	609
50553323-38f6-4042-a9f6-f953e7fe9072	Nones et al 2014	331	1	0	0	0	10	78	0	417
52b05e15-0412-4679-92f2-0b97cc69d4f4	Nones et al 2014	359	2	0	0	0	12	47	0	417
5871129b-4b66-4261-b326-353c41aa4cce	Nones et al 2014	112	5	0	0	0	3	21	0	140
5fd4044e-2b5b-4747-9ec3-5022830abd63	Nones et al 2014	117	0	0	0	0	13	72	0	202
6804574e-a38a-48f8-ab1a-bbfd12c524a9	Nones et al 2014	490	2	0	0	0	20	92	0	602
6e394b7c-8555-4a64-b342-e1d952991f29	Nones et al 2014	169	59	0	0	0	8	75	0	310
6ed62b87-fd7d-4c51-8a7b-2cae3fea56de	Nones et al 2014	575	37	0	0	0	21	100	0	723
76d63ac1-5985-47b8-9236-0f821572bfaf	Nones et al 2014	473	18	0	0	0	12	79	0	574
7deb8b86-e212-41ca-9d14-de0007b7ac90	Nones et al 2014	437	1	0	0	0	15	50	0	499
8ba60960-727c-4f82-b545-10087b99eb7c	Nones et al 2014	105	36	0	0	0	8	33	0	182
9bb9ff43-7a94-4823-8d4d-97e1d383fe3b	Nones et al 2014	575	2	0	0	0	9	65	0	645
aade4a47-aaeb-40ce-ab3b-cba53c34cc41	Nones et al 2014	425	47	0	0	0	5	39	0	516
bb9aad8a-7462-4797-90b7-b35d9aed287c	Nones et al 2014	22	40	0	0	0	7	76	0	145

dbda7a52-9043-4510-9145-b2c998dd3d91	Nones et al 2014	203	1	0	0	0	2	20	0	226
e03d78a2-3481-4c89-868b-b8c2f9148445	Nones et al 2014	243	0	0	0	0	9	77	0	326
e5789831-aee2-43f2-910c-605b04705c16	Nones et al 2014	40	22	0	0	0	7	38	0	107
fe1ac755-51c4-4735-b94a-f4a5d1279986	Nones et al 2014	259	2	0	0	0	10	39	0	310
TCGA-2H-A9GF-01	TCGA	3504	41	0	0	0	18	156	0	3681
TCGA-2H-A9GH-01	TCGA	2145	0	0	0	0	9	74	0	2215
TCGA-2H-A9GI-01	TCGA	1737	0	0	0	0	23	120	1	1856
TCGA-2H-A9GJ-01	TCGA	3332	0	0	0	0	8	82	0	3404
TCGA-2H-A9GK-01	TCGA	986	0	0	0	0	12	119	0	1112
TCGA-2H-A9GL-01	TCGA	1153	0	0	0	0	16	109	0	1265
TCGA-2H-A9GM-01	TCGA	1748	16	0	0	0	9	79	0	1842
TCGA-2H-A9GN-01	TCGA	3958	3	0	0	0	8	69	0	4016
TCGA-2H-A9GO-01	TCGA	1450	0	0	0	0	8	81	0	1535
TCGA-2H-A9GQ-01	TCGA	2596	15	0	0	0	13	86	0	2695
TCGA-2H-A9GR-01	TCGA	696	0	0	0	0	17	170	0	876
TCGA-IC-A6RE-01	TCGA	1163	0	0	0	0	24	259	1	1433
TCGA-IG-A4QS-01	TCGA	3005	0	0	0	0	20	111	0	3111
TCGA-IG-A7DP-01	TCGA	0	0	0	0	0	1	16	0	17
TCGA-JY-A6F8-01	TCGA	1712	1	0	0	0	15	99	0	1816
TCGA-JY-A6FB-01	TCGA	1284	0	0	0	0	14	75	0	1361
TCGA-JY-A6FH-01	TCGA	1661	1	0	0	0	7	87	0	1744
TCGA-JY-A938-01	TCGA	1290	0	0	0	0	10	73	0	1365
TCGA-JY-A939-01	TCGA	170	0	0	0	0	2	68	0	240
TCGA-JY-A93C-01	TCGA	1417	0	0	0	0	14	53	0	1478
TCGA-JY-A93D-01	TCGA	503	0	0	0	0	9	90	0	600
TCGA-JY-A93E-01	TCGA	1086	0	0	0	0	13	120	0	1208
TCGA-L5-A43C-01	TCGA	333	0	0	0	0	9	66	0	407
TCGA-L5-A43E-01	TCGA	1371	1	0	0	0	9	83	0	1451
TCGA-L5-A43I-01	TCGA	2665	0	0	0	0	15	89	0	2758
TCGA-L5-A43M-01	TCGA	805	0	0	0	0	2	28	0	833
TCGA-L5-A4OE-01	TCGA	2230	0	0	0	0	33	149	0	2392
TCGA-L5-A4OF-01	TCGA	2153	0	0	0	0	8	43	0	2197
TCGA-L5-A4OG-01	TCGA	1167	0	0	0	0	15	72	0	1240
TCGA-L5-A4OH-01	TCGA	4110	0	0	0	0	10	155	0	4222
TCGA-L5-A4OJ-01	TCGA	593	1	0	0	0	25	159	0	766
TCGA-L5-A4ON-01	TCGA	3130	0	0	0	0	8	76	0	3195
TCGA-L5-A4OO-01	TCGA	460	0	0	0	0	10	48	0	517
TCGA-L5-A4OP-01	TCGA	1476	0	0	0	0	5	56	0	1531
TCGA-L5-A4OQ-01	TCGA	2227	0	0	0	0	6	29	0	2259

TCGA-L5-A4OR-01	TCGA	2225	1	0	0	0	16	85	1	2309
TCGA-L5-A4OS-01	TCGA	875	0	0	0	0	5	41	0	917
TCGA-L5-A4OT-01	TCGA	1521	0	0	0	0	12	78	0	1600
TCGA-L5-A4OU-01	TCGA	1115	0	0	0	0	11	75	0	1199
TCGA-L5-A4OW-01	TCGA	1882	0	0	0	0	10	104	0	1984
TCGA-L5-A4OX-01	TCGA	1364	1	0	0	0	12	57	0	1429
TCGA-L5-A88T-01	TCGA	25	0	0	0	0	4	40	0	68
TCGA-L5-A88V-01	TCGA	2453	1	0	0	0	17	84	0	2536
TCGA-L5-A88Y-01	TCGA	1812	0	0	0	0	14	90	0	1903
TCGA-L5-A891-01	TCGA	3916	0	0	0	0	17	110	0	4010
TCGA-L5-A893-01	TCGA	1220	0	0	0	0	10	94	0	1316
TCGA-L5-A8NE-01	TCGA	2795	1	0	0	0	17	116	0	2913
TCGA-L5-A8NF-01	TCGA	2945	2	0	0	0	18	77	0	3020
TCGA-L5-A8NG-01	TCGA	831	0	0	0	0	8	98	0	934
TCGA-L5-A8NH-01	TCGA	2406	0	0	0	0	9	98	0	2499
TCGA-L5-A8NI-01	TCGA	1918	0	0	0	0	11	115	0	2020
TCGA-L5-A8NJ-01	TCGA	3462	4	0	0	0	12	125	0	3573
TCGA-L5-A8NL-01	TCGA	593	0	0	0	0	9	96	0	694
TCGA-L5-A8NN-01	TCGA	2930	0	0	0	0	11	76	0	3007
TCGA-L5-A8NR-01	TCGA	2122	0	0	0	0	22	98	0	2222
TCGA-L5-A8NS-01	TCGA	1032	0	0	0	0	28	191	0	1235
TCGA-L5-A8NT-01	TCGA	1251	0	0	0	0	10	87	0	1339
TCGA-L5-A8NU-01	TCGA	42	0	0	0	0	5	30	0	77
TCGA-L5-A8NV-01	TCGA	782	0	0	0	0	8	89	0	876
TCGA-L5-A8NW-01	TCGA	3588	0	0	0	0	21	89	0	3678
TCGA-L7-A6VZ-01	TCGA	2353	1	0	0	0	9	129	0	2477
TCGA-M9-A5M8-01	TCGA	679	0	0	0	0	9	34	0	720
TCGA-Q9-A6FW-01	TCGA	789	0	0	0	0	10	97	0	892
TCGA-R6-A6DN-01	TCGA	3318	2	0	0	0	9	50	0	3365
TCGA-R6-A6DQ-01	TCGA	2112	0	0	0	0	6	49	1	2164
TCGA-R6-A6KZ-01	TCGA	1884	0	0	0	0	6	79	0	1963
TCGA-R6-A6L4-01	TCGA	4398	0	0	0	0	7	59	0	4447
TCGA-R6-A6L6-01	TCGA	772	0	0	0	0	13	75	0	854
TCGA-R6-A6XG-01	TCGA	2030	49	0	0	0	12	132	0	2211
TCGA-R6-A6XQ-01	TCGA	1873	11	0	0	0	11	77	0	1963
TCGA-R6-A6Y0-01	TCGA	1216	0	0	0	0	16	113	0	1338
TCGA-R6-A6Y2-01	TCGA	773	0	0	0	0	14	90	0	873
TCGA-R6-A8W5-01	TCGA	1037	3	0	0	0	10	56	0	1105
TCGA-R6-A8W8-01	TCGA	2447	1	0	0	0	11	69	1	2517

TCGA-R6-A8WC-01	TCGA	848	2	0	0	0	16	90	0	951
TCGA-R6-A8WG-01	TCGA	1805	1	0	0	0	8	85	0	1892
TCGA-RE-A7BO-01	TCGA	2225	0	0	0	0	11	137	0	2348
TCGA-S8-A6BV-01	TCGA	1146	0	0	0	0	9	76	0	1223
TCGA-V5-A7RB-01	TCGA	1677	0	0	0	0	12	105	0	1783
TCGA-V5-A7RE-01	TCGA	1734	0	0	0	0	11	77	0	1805
TCGA-V5-AASW-01	TCGA	3598	4	0	0	0	7	68	1	3664
TCGA-V5-AASX-01	TCGA	771	0	0	0	0	16	155	0	933
TCGA-VR-A8EQ-01	TCGA	1193	0	0	0	0	21	106	0	1308
TCGA-VR-AA4D-01	TCGA	1319	0	0	0	0	9	61	0	1385
TCGA-X8-AAAR-01	TCGA	439	0	0	0	0	6	71	0	514
TCGA-ZR-A9CJ-01	TCGA	1118	0	0	0	0	12	69	0	1195

Table 3.8. Most recurrently altered cancer helpers in the 261 OACs. For each gene reported are the gene description, the number and percentage of samples where it is altered and the type of alteration.

				Type of alteration (n of samples)							
Gene	**Description**	**Total altered samples (n)**	**Total altered samples (%)**	**Gene gain**	**Gene loss**	**Gene translocation**	**Gene inversion**	**Gene insertion**	**Truncating**	**Non-truncating damaging**	**Gain of function**
TOMM34	translocase of outer mitochondrial membrane 34	33	12.6	33	0	0	0	0	0	0	0
NCOA3	nuclear receptor coactivator 3	32	12.3	30	0	0	1	0	0	1	0
E2F1	E2F transcription factor 1	29	11.1	29	0	0	0	0	0	0	0
MCM7	minichromosome maintenance complex component 7	28	10.7	26	0	0	0	0	0	2	0
VAPB	VAMP (vesicle-associated membrane protein)-associated protein B and C	26	10	26	0	0	0	0	0	0	0
DNMT3B	DNA (cytosine-5-)-methyltransferase 3 beta	25	9.6	24	0	0	0	0	1	0	0
BAG6	BCL2-associated athanogene 6	24	9.2	23	0	0	0	0	0	1	0
DLC1	DLC1 Rho GTPase activating protein	24	9.2	19	0	5	8	0	0	6	0
VEGFA	vascular endothelial growth factor A	23	8.8	23	0	0	0	0	0	0	0
ASAP1	ArfGAP with SH3 domain, ankyrin repeat and PH domain 1	20	7.7	20	0	0	1	0	0	0	0

THRA	thyroid hormone receptor, alpha	20	7.7	20	0	1	2	0	0	0	0
PRRC2A	proline-rich coiled-coil 2A	19	7.3	18	0	0	2	0	0	2	0
SCRIB	scribbled planar cell polarity protein	19	7.3	19	0	0	0	0	0	0	0
LONRF1	LON peptidase N-terminal domain and ring finger 1	18	6.9	18	0	0	3	0	0	0	0
PLOD3	procollagen-lysine, 2-oxoglutarate 5-dioxygenase 3	18	6.9	18	0	0	0	0	0	0	0
PTPN1	protein tyrosine phosphatase, non-receptor type 1	18	6.9	18	0	0	0	0	0	0	0
NBEA	neurobeachin	17	6.5	9	0	1	0	0	0	10	0
AGO2	argonaute RISC catalytic component 2	16	6.1	15	0	0	0	0	0	1	0
PAK4	p21 protein (Cdc42/Rac)-activated kinase 4	15	5.7	14	0	0	1	0	0	1	0
ROCK1	Rho-associated, coiled-coil containing protein kinase 1	15	5.7	13	0	0	6	0	0	0	0
ACTN4	actinin, alpha 4	14	5.4	13	0	0	1	0	0	0	0
HSPH1	heat shock 105kDa/110kDa protein 1	14	5.4	14	0	0	0	0	0	0	0
JUP	junction plakoglobin	14	5.4	12	0	3	6	0	0	0	0
PTK2	protein tyrosine kinase 2	14	5.4	13	0	0	0	0	0	2	0
SOX5	SRY (sex determining region Y)-box 5	14	5.4	10	0	1	4	0	0	5	0
STK3	serine/threonine kinase 3	14	5.4	14	0	0	0	0	0	0	0

3.3.7 Patient-specific helpers perturb related biological processes

The observation that OAC helpers are mostly rare or patient-specific (Figure 3.11C) poses the question of whether these genes act on similar or different biological processes. Therefore, I sought to investigate the pathways in which helpers are enriched in to gain insight into their cancer promoting functions. To this end, I analysed the biological processes perturbed by helpers and compared them to those of drivers, as a means to quantify both the novel and known biological processes that helpers contribute to.

I manually reviewed all 476 known cancer genes (Forbes et al. 2017) with damaging alterations in the OCCAMS cohort and retained 202 of them based on the concordance between the type of acquired modification and the literature evidence of their cancer role (Methods, summarized in Appendix Table 7.1). Analysis of these 202 genes showed that the median number of drivers per OAC was seven (Figure 3.13A) and their recurrence profile was skewed towards high values, with more than 25% of genes altered in 10 samples or more (Figure 3.13B). Both of these characteristics were in accordance with recent estimates (Martincorena et al. 2017; Sabarinathan et al. 2017). Finally, as expected from the high number of genomic regions in OAC undergoing amplifications, the majority (nearly 50%) of these drivers were subject to somatic copy number amplifications (Figure 3.13C).

I then performed two independent gene set enrichment analyses[4], one with the 202 known drivers and one with the 952 helpers, to dissect their relative functional contribution to OAC. This led to 212 and 189 enriched pathways out of the 1,877 tested in drivers and helpers, respectively (Appendix Tables 7.3 and 7.4). Interestingly, the analysis of known drivers resulted in a

[4] Analysis was performed in collaboration with Joel Nulsen

higher number of enriched pathways than helpers, despite their lower number. This reflected the higher number of pathways that drivers mapped to (median of four pathways for known drivers and two pathways for helpers). As expected, owing to the high alteration recurrence of known drivers, OACs had on average more enriched pathways due to the contribution of known drivers rather than helpers (Figure 3.13D, E).

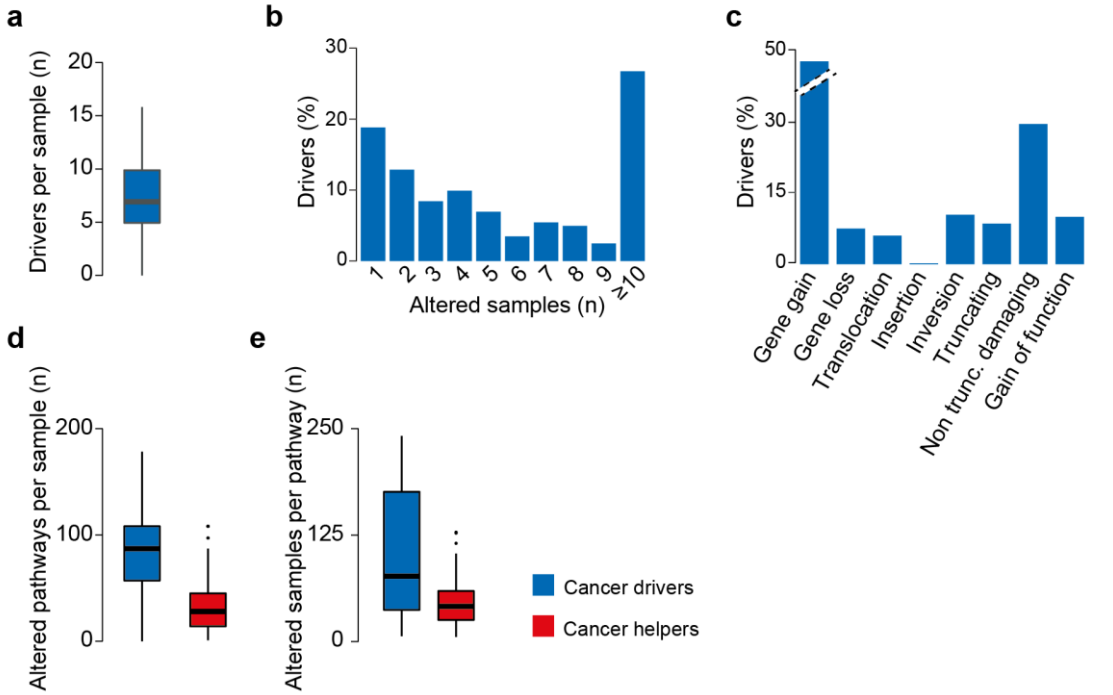


Figure 3.13. Characteristics of cancer drivers. **(a)** Distribution of known drivers across 261 OACs. **(b)** Recurrence of cancer drivers across 261 OACs. Only samples acquiring alterations with a damaging effect are considered. **(c)** Distribution of damaging alterations in 202 cancer drivers. Overall, these genes acquire 1,967 damaging alterations. Distribution of altered pathways per samples (**d**) and altered samples per pathway (**e**) for known drivers and newly predicted helpers.

Seventy-three pathways (34%) enriched in known drivers were perturbed in more than 50% of OACs (Appendix Table 7.3, Figure 3.14). These ‘universal

cancer pathways' are involved in well-known cancer-related processes[5], such as intracellular signalling, cell cycle control, apoptosis and DNA repair, and are associated with the most recurrently altered known drivers (*TP53, CDKN2A, MYC, ERBB2, SMAD4, CDK6, KRAS*). Interestingly, 51 of the 73 (70%) were also enriched in helpers and 86 patients with altered helpers in a universal cancer pathway had no known drivers in that pathway (Figure 3.15). This indicates that helpers often contribute to the perturbation of key cancer pathways and that their alteration may be sufficient for cancer development in the absence of known drivers.

Taken together, these results demonstrated that helpers contribute towards the perturbation of well-known cancer-related pathways. The implication of helpers in these pathways denotes that these genes are true positive predictions, as they are direct interactors of known drivers. However, further experimental validation is needed in order to dissect their role in tumorigenesis (see chapter 4). Finally, the high number of OACs (n=86) with helpers, but no known driver in those pathways, suggests that helpers contribute towards the refinement of the tumorigenic molecular landscape of these samples. Therefore, both helpers and known drivers should be considered when tumorigenic alterations are taken into account.

[5] Analysis of 'universal cancer pathways' was performed in collaboration with Dr. Elizabeth Foxall

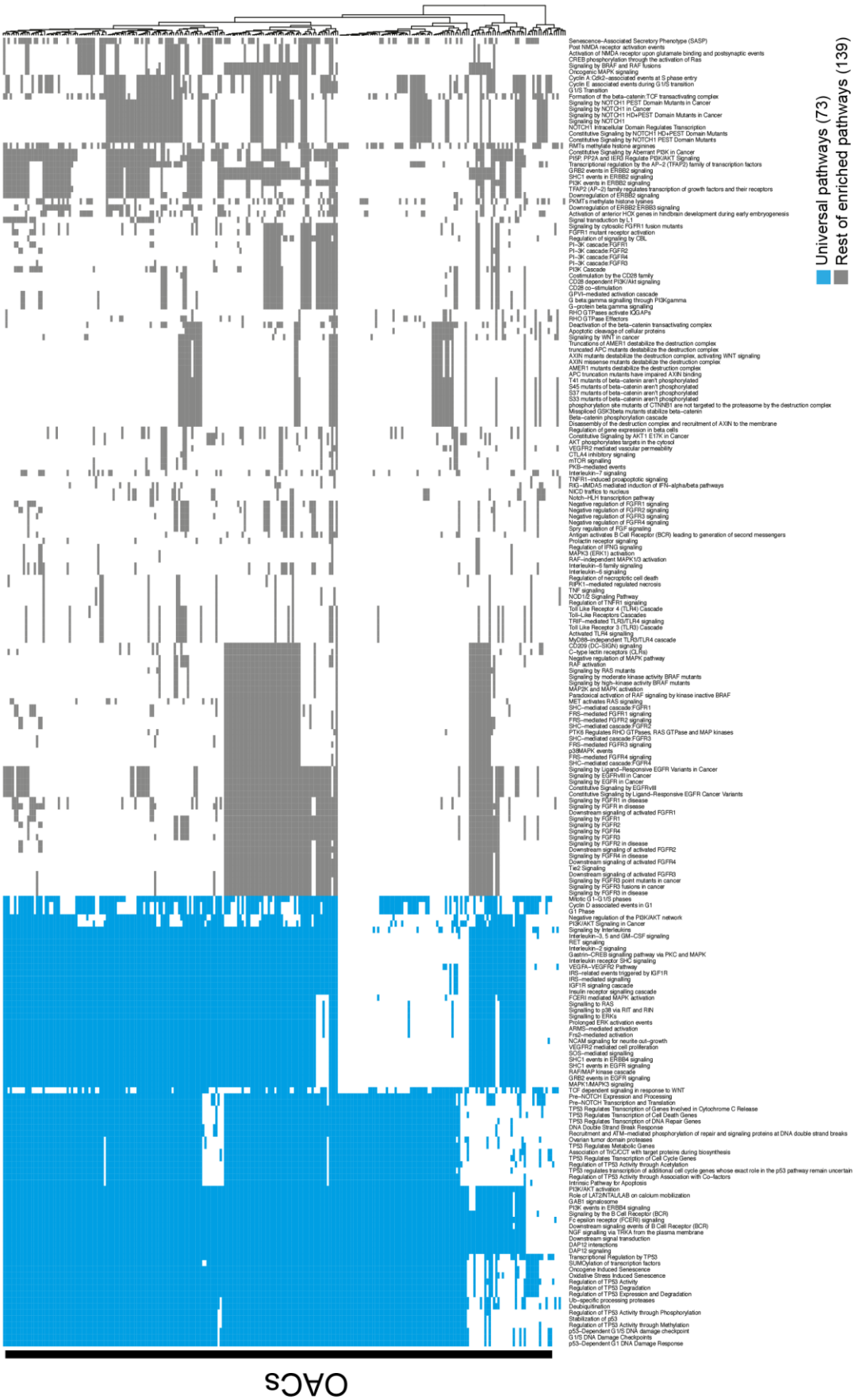

Universal pathways (73)
Rest of enriched pathways (139)
OACS
Senescence−Associated Secretory Phenotype (SASP)
Post NMDA receptor activation events
Activation of NMDA receptor upon glutamate binding and postsynaptic events
CREB phosphorylation through the activation of Ras
Signaling by BRAF and RAF fusions
Oncogenic MAPK signaling
Cyclin A:Cdk2−associated events at S phase entry
Cyclin E associated events during G1/S transition
G1/S Transition
Formation of the beta−catenin:TCF transactivating complex
Signaling by NOTCH1 PEST Domain Mutants in Cancer
Signaling by NOTCH1 in Cancer
Signaling by NOTCH1 HD+PEST Domain Mutants in Cancer
Signaling by NOTCH1
NOTCH1 Intracellular Domain Regulates Transcription
Constitutive Signaling by NOTCH1 HD+PEST Domain Mutants
Constitutive Signaling by NOTCH1 PEST Domain Mutants
RMTs methylate histone arginines
Constitutive Signaling by Aberrant PI3K in Cancer
PIP3 PP2A and IER3 Regulate PI3K/AKT Signaling
Transcriptional regulation by the AP−2 (TFAP2) family of transcription factors
GRB2 events in ERBB2 signaling
SHC1 events in ERBB2 signaling
PI3K events in ERBB2 signaling
TFAP2 (AP−2) family regulates transcription of growth factors and their receptors
Downregulation of ERBB2 signaling
PKMTs methylate histone lysines
Downregulation of ERBB2:ERBB3 signaling
Activation of anterior HOX genes in hindbrain development during early embryogenesis
Signal transduction by L1
Signaling by cytosolic FGFR1 fusion mutants
FGFR1 mutant receptor activation
Regulation of signaling by CBL
PI−3K cascade:FGFR1
PI−3K cascade:FGFR2
PI−3K cascade:FGFR4
PI−3K cascade:FGFR3
PI3K Cascade
Costimulation by the CD28 family
CD28 dependent PI3K/Akt signaling
CD28 co−stimulation
GPVI−mediated activation cascade
G beta:gamma signalling through PI3Kgamma
G−protein beta:gamma signalling
RHO GTPases activate IQGAPs
RHO GTPase Effectors
Deactivation of the beta−catenin transactivating complex
Apoptotic cleavage of cellular proteins
Signaling by WNT in cancer
Truncations of AMER1 destabilize the destruction complex
truncated APC mutants destabilize the destruction complex
AXIN mutants destabilize the destruction complex, activating WNT signaling
AXIN missense mutants destabilize the destruction complex
AMER1 mutants destabilize the destruction complex
APC truncation mutants have impaired AXIN binding
T41 mutants of beta−catenin aren't phosphorylated
S45 mutants of beta−catenin aren't phosphorylated
S37 mutants of beta−catenin aren't phosphorylated
S33 mutants of beta−catenin aren't phosphorylated
phosphorylation site mutants of CTNNB1 are not targeted to the proteasome by the destruction complex
Misspliced GSK3beta mutants stabilize beta−catenin
Beta−catenin phosphorylation cascade
Disassembly of the destruction complex and recruitment of AXIN to the membrane
Regulation of gene expression in beta cells
Constitutive Signaling by AKT1 E17K in Cancer
AKT phosphorylates targets in the cytosol
VEGFR2 mediated vascular permeability
CTLA4 inhibitory signaling
mTOR signalling
PKB−mediated events
Interleukin−7 signaling
TNFR1−induced proapoptotic signaling
RIG−I/MDA5 mediated induction of IFN−alpha/beta pathways
NICD traffics to nucleus
Notch−HLH transcription pathway
Negative regulation of FGFR1 signaling
Negative regulation of FGFR2 signaling
Negative regulation of FGFR3 signaling
Negative regulation of FGFR4 signaling
Spry regulation of FGF signaling
Antigen activates B Cell Receptor (BCR) leading to generation of second messengers
Prolactin receptor signaling
Regulation of IFNG signaling
MAPK3 (ERK1) activation
RAF−independent MAPK1/3 activation
Interleukin−6 family signaling
Interleukin−6 signaling
Regulation of necroptotic cell death
RIPK1−mediated regulated necrosis
TNF signaling
NOD1/2 Signaling Pathway
Regulation of TNFR1 signaling
Toll Like Receptor 4 (TLR4) Cascade
Toll−Like Receptors Cascades
TRIF−mediated TLR3/TLR4 signaling
Toll Like Receptor 3 (TLR3) Cascade
Activated TLR4 signalling
MyD88−independent TLR3/TLR4 cascade
CD209 (DC−SIGN) signaling
C−type lectin receptors (CLRs)
Negative regulation of MAPK pathway
RAF activation
Signaling by RAS mutants
Signaling by moderate kinase activity BRAF mutants
Signaling by high−kinase activity BRAF mutants
MAPK and MAPK activation
Paradoxical activation of RAF signaling by kinase inactive BRAF
MET activates RAS signaling
SHC−mediated cascade:FGFR1
FRS−mediated FGFR1 signaling
FRS−mediated FGFR2 signaling
SHC−mediated cascade:FGFR2
PTK6 Regulates RHO GTPases, RAS GTPase and MAP kinases
SHC−mediated cascade:FGFR3
FRS−mediated FGFR3 signaling
p38MAPK events
FRS−mediated FGFR4 signaling
SHC−mediated cascade:FGFR4
Signaling by Ligand−Responsive EGFR Variants in Cancer
Signaling by EGFRvIII in Cancer
Signaling by EGFR in Cancer
Constitutive Signaling by EGFRvIII
Constitutive Signaling by Ligand−Responsive EGFR Cancer Variants
Signaling by FGFR1 in disease
Signaling by FGFR in disease
Downstream signaling of activated FGFR1
Signaling by FGFR1
Signaling by FGFR2
Signaling by FGFR4
Signaling by FGFR3
Signaling by FGFR2 in disease
Downstream signaling of activated FGFR2
Signaling by FGFR4 in disease
Downstream signaling of activated FGFR4
Tie2 Signaling
Downstream signaling of activated FGFR3
Signaling by FGFR3 point mutants in cancer
Signaling by FGFR3 fusions in cancer
Signaling by FGFR3 in disease
Mitotic G1−G1/S phases
Cyclin D associated events in G1
G1 Phase
Negative regulation of the PI3K/AKT network
PI3K/AKT Signaling in Cancer
Signaling by Interleukins
Interleukin−3, 5 and GM−CSF signaling
RET signaling
Interleukin−2 signaling
Gastrin−CREB signalling pathway via PKC and MAPK
Interleukin receptor SHC signaling
VEGFR−VEGFR2 Pathway
IRS−related events triggered by IGF1R
IRS−mediated signalling
IGF1R signaling cascade
Insulin receptor signalling cascade
FCERI mediated MAPK activation
Signalling to RAS
Signalling to p38 via RIT and RIN
Signalling to ERKs
Prolonged ERK activation events
ARMS−mediated activation
Frs2−mediated activation
NCAM signaling for neurite out−growth
VEGFR2 mediated cell proliferation
SOS−mediated signalling
SHC1 events in ERBB4 signaling
SHC1 events in EGFR signaling
RAF/MAP kinase cascade
GRB2 events in EGFR signaling
MAPK1/MAPK3 signaling
TCF dependent signaling in response to WNT
Pre−NOTCH Expression and Processing
Pre−NOTCH Transcription and Translation
TP53 Regulates Transcription of Genes Involved in Cytochrome C Release
TP53 Regulates Transcription of Cell Death Genes
TP53 Regulates Transcription of DNA Repair Genes
DNA Double Strand Break Response
Recruitment and ATM−mediated phosphorylation of repair and signaling proteins at DNA double strand breaks
Ovarian tumor domain proteases
TP53 Regulates Metabolic Genes
Association of TriC/CCT with target proteins during biosynthesis
TP53 Regulates Transcription of Cell Cycle Genes
Regulation of TP53 Activity through Acetylation
TP53 regulates transcription of additional cell cycle genes whose exact role in the p53 pathway remain uncertain
Regulation of TP53 Activity through Association with Co−factors
Intrinsic Pathway for Apoptosis
PI3K/AKT activation
Role of LAT2/NTAL/LAB on calcium mobilization
GAB1 signalosome
PI3K events in ERBB4 signaling
Signaling by the B Cell Receptor (BCR)
Fc epsilon receptor (FCERI) signaling
Downstream signaling events of B Cell Receptor (BCR)
NGF signalling via TRKA from the plasma membrane
Downstream signal transduction
DAP12 interactions
DAP12 signaling
Transcriptional Regulation by TP53
SUMOylation of transcription factors
Oncogene Induced Senescence
Oxidative Stress Induced Senescence
Regulation of TP53 Activity
Regulation of TP53 Degradation
Regulation of TP53 Expression and Degradation
Ub−specific processing proteases
Deubiquitination
Regulation of TP53 Activity through Phosphorylation
Stabilization of p53
Regulation of TP53 Activity through Methylation
p53−Dependent G1/S DNA damage checkpoint
G1/S DNA Damage Checkpoints
p53−Dependent G1 DNA Damage Response

Figure 3.14. Hierarchical clustering on the presence/absence matrix of samples and perturbed pathways was performed as described in Methods. Each row represents a sample and each column an enriched pathway. Samples were assigned to a given pathway if they had at least one altered known driver mapping to that pathway. Seventy-three universal pathways perturbed in at least 50% of samples are coloured in light blue.

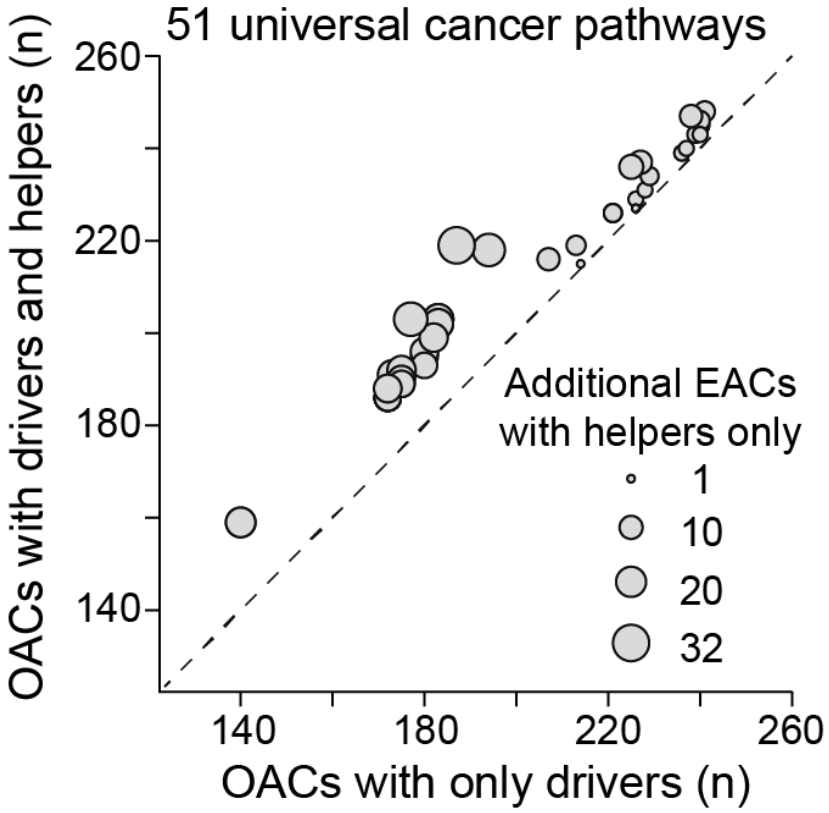


Figure 3.15. Scatterplot of 51 ‘universal’ pathways enriched in known drivers and helpers. For each pathway, the number of OACs with altered drivers and the number of OACs with altered drivers and helpers is shown. The size of dots is proportional to the additional OACs with perturbations in these pathways because of altered helpers only.

3.3.8 Mutational signatures in OAC helpers

Different mutational processes generate specific patterns of point mutations in OAC genomes. Some of these processes have been previously described (Nik-Zainal et al. 2012; Alexandrov et al. 2013), but several signatures still remain of unknown origin. In OAC, five main signatures have been reported previously (Secrier et al. 2016) and can be found summarised in COSMIC (https://cancer.sanger.ac.uk/cosmic/signatures):

- S1 (endogenous mutational process initiated by spontaneous deamination of 5-methylcytosine),
- S2 (attributed to activity of the AID/APOBEC family of cytidine deaminases),
- S3 (associated with *BRCA1* and *BRCA2* mutations),
- S17 (of unknown aetiology, but associated with “mutator” phenotype in OAC),
- S18 (of unknown aetiology with strong preference for C>A mutations in a GCA and TCT trinucleotide context) and their prevalence within OAC genomes facilitated patient stratification with potential therapeutic implications.

To investigate the mutational processes operating on cancer helpers, I analysed all point mutations that were associated with them. To this end, I compiled a list of 5,898 somatic substitutions corresponding to an average of 23.4 mutations per OAC (ranging from 1 to 270). In total, 251 OACs were included in this analysis as 10 OACs had only amplified helpers. Overall, three out of five previously described signatures were identified in helpers (Figure 3.16A) and those were: signature 1 (cosine similarity of 0.79 with signature 1 from COSMIC), signature 2 (cosine similarity of 0.83 with signature 2 from

COSMIC) and signature 17 (cosine similarity of 0.98 with signature 17 from COSMIC). No new signatures were identified in this analysis. Confirming previous results, there was an association between the number of mutations in helpers of each OAC and the exposure to signature 17 (characteristic of the mutagenic group in Secrier et al.) (Figure 3.16B). Finally, this analysis further confirmed that mutations in helpers are not sequencing artefacts, as no atypical mutational signatures were retrieved.

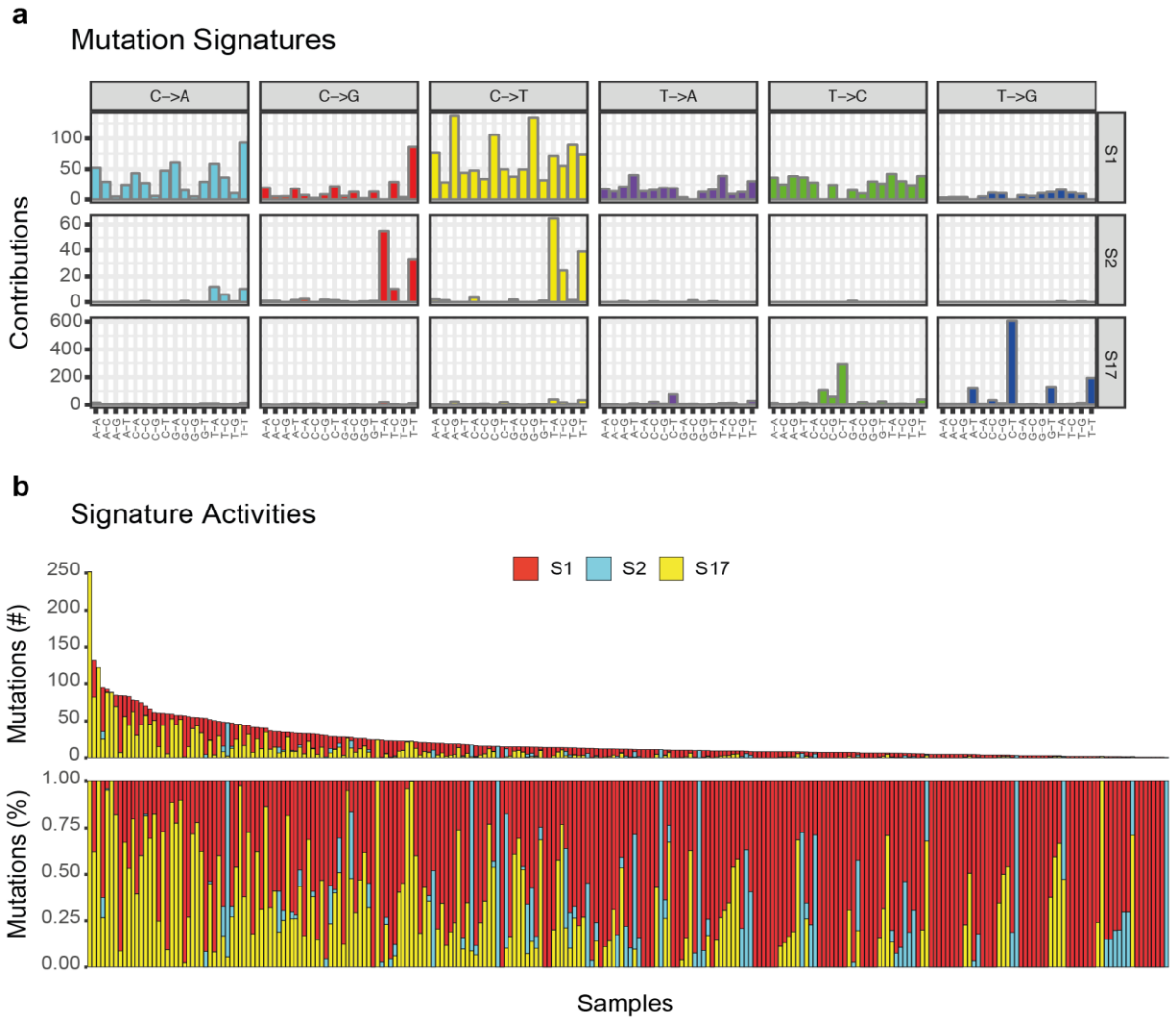

a
Mutation Signatures
C->A
C->G
C->T
T->A
T->C
T->G
S1
S2
S17
Contributions
b
Signature Activities
S1
S2
S17
Mutations (#)
Mutations (%)
Samples

Figure 3.16. Mutational signature analysis of point mutations in helpers. **(a)** Three mutational signatures discovered in 251 cohort with point mutations in helpers. Analysis was performed using BayesNMF (Tan and Fevotte 2013). **(b)** Exposure of each OAC to the mutational signature discovered in a. Exposure is plotted as mutation count (upper panel) and fraction (bottom panel).

3.4 Discussion

In this chapter, I described the application of sysSVM to identify rare and patient-specific driver genes in OAC, a genomically unstable cancer type. High genomic instability leads to inter-patient heterogeneity and a high number of genes being somatically altered in low number of OACs. The working hypothesis of this study was that, in the absence of recurrently mutated genes, major driver genes are complemented by multiple privately (or rarely) altered genes that contribute to tumorigenesis (named cancer helpers). Although sysSVM was not developed to identify exclusively cancer helpers, but cancer drivers in general, its ability to score genes based on their driver potential in individual samples renders OAC a very good case study for the application of sysSVM. Of note, sysSVM could successfully identify known drivers, when applied to previously unseen OAC cohorts (Figure 3.10).

Cancer helpers are expected to be subject to weak selection. In support of this hypothesis, the overall selection acting on oesophageal cancer genomes is among the lowest across cancer types (Martincorena et al. 2017), despite a median of 382 damaged genes per OAC. This indicates that the exclusive focus on genes under strong selection is likely to return only a partial representation of the genes as important contributors to OAC.

An advantage of not measuring selection directly is that sysSVM considers all types of alterations (SNVs, indels, CNVs, and structural variations) simultaneously. These alterations combined with the systems-level properties, which have been shown to be discriminant of known cancer genes from the rest of human genes, can be used to derive a global similarity metric of any altered gene to known cancer drivers. Therefore, a comprehensive overview of the genetic modifications that play a cancer promoting role in each patient can be constructed regardless of how frequent each alteration is.

I applied sysSVM to 261 OACs from ICGC and prioritized 952 genes that, together with known drivers, help cancer progression. The large number of these cancer helpers is in agreement with the recent observation of a positive correlation between mutational burden and number of driver genes, which is only partially explained by a sample size effect (Bailey et al. 2018). As speculated in the introduction of this chapter, it is possible that this positive correlation may indicate that the number of functionally relevant genes increases with the number of altered genes. However, this might not be true for all cancer types and further research is warranted towards deciphering the relationship between the number of alterations and the number of drivers and helpers in a pan-cancer setting.

As a further proof towards the existence of cancer helpers, I showed in this chapter that helpers, indeed, converge towards the perturbation of the same pathways. In other words, although helpers are possibly subject to weak selection, the perturbation of the pathways that they are involved in are important for tumorigenesis in OAC. Several of the processes that were discovered to be enriched here, due to helpers, are well-known contributors to cancer development, and include intracellular signalling, cell cycle control, and DNA repair. Interestingly,

while the known drivers tend to encode upstream players in these processes, helpers are often downstream effectors, suggesting a more local role of helpers at a single patient level.

The interplay of cancer helpers and major drivers is not fully understood. Recent mathematical modelling of the complex relationship between major drivers and beneficial passengers (or cancer helpers as those defined in this thesis) showed that tumour cell population grows continuously with varying growth rates (Li and Thirumalai 2016). When deleterious passengers were also considered in the models, population homeostasis was observed, which can be supported by the observations that tumour growth is often interrupted by periods of dormancy (Ghajar et al. 2013). I anticipate that the function of cancer helpers is context-dependant and their tumorigenic potential will be dependent on not only the combination of major drivers, but also the existence of other cancer helpers. This makes our ability to experimentally validate these genes quite challenging, as the readouts of the experimental assays also need to be context-dependant.

In summary, I described one of the first attempts to extend the discovery of acquired perturbations contributing to cancer, beyond those of recurrent drivers. An obvious criticism to this work is that it depends on the current knowledge of known cancer genes, as it needs a well-defined training set. However, I argue that the discovery rate of known (and recurrent) cancer genes has already reached a plateau, therefore our capacity to identify new recurrently mutated genes is limited. Moreover, a machine learning framework, such as that of sysSVM, can provide a useful tool, whose models will become increasingly better as new cancer genes are discovered. Therefore, I envision the use of sysSVM as a dynamic feedback

process, during which newly-identified and validated drivers and helpers will transition from prediction to training set, allowing the discovery of more novel driver genes. Additional efforts are warranted to fully exploit the potential of these approaches with the aim to offer a more comprehensive view of the molecular mechanisms behind cancer and to guide novel clinical interventions.

Chapter 4. Oesophageal cancer patient stratification using sysSVM predictions

4.1 Chapter overview

In this chapter, I describe the utility of sysSVM predictions to stratify patients into subgroups in an effort to understand the biology of OAC and suggest putative therapeutic interventions. After showing that cancer helpers converge towards perturbations of similar pathways (chapter 3), I now use the perturbed biological processes instead of individual helper genes to define patient groups. Previous approaches for patient stratification using molecular data proved informative for clinical practice in multiple cancer types, including OAC, in which a proportion of patients (~20%) responded positively to trastuzumab, a monoclonal antibody targeting HER2 receptor tyrosine kinase (Bang et al. 2010). Most recent studies focused on mutational signatures and defined three distinct molecular subtypes in OAC with potential therapeutic relevance (Secrier et al. 2016). Therefore, additional studies are needed to dissect the full potential of molecular subtyping of OAC patients and inform precision medicine. This part of my thesis builds on the potential of a well-described compendium of cancer drivers and its utility to stratify patients in OAC.

4.2 Introduction

Despite the advances in drug development, not all patients respond favourably to treatment. Many available therapies, such as chemotherapy and immunotherapy, have not led to consistently high cure rates in many cancer types

(Crawford 2013; Emens et al. 2017). However, the fraction of patients that does not respond to cancer therapy varies significantly across different cancer types, with certain types being more amenable to treatment. For instance, current treatment protocols for most childhood hematologic malignancies exhibit high success rates (Suttorp et al. 2018); five-year survival rates of patients with acute lymphoblastic leukaemia (ALL), the commonest childhood cancer, are as high as 90% (Saletta, Seng, and Lau 2014). In contrast, in adult advanced melanoma only 20% of patients respond to immunotherapy treatments, such as ipilimumab (monoclonal antibody targeting CTLA-4 on the surface of T-cells) monotherapy (Schadendorf et al. 2015).

A possible explanation to these differences across cancer types might be the fact that, often, treatment strategies are chosen according to clinical and pathological criteria, ignoring molecular alterations and subgroups of patients. Historically, clinical management of cancer has been guided by the organ system classification, primarily relying on histological and immunohistochemical analysis of tumour tissues (Song, Merajver, and Li 2015). In the past decade, this approach has begun to change, and new tumour stratification systems based on specific molecular alterations have started to emerge (Torkamani, Verkhivker, and Schork 2009). Many cancer genome sequencing initiatives, such as TCGA and ICGC, contributed to this change, by unravelling the somatic alterations in cancer genomes (Lander and Weinberg 2000; Kaiser 2008; Collins and Barker 2007). Analysis of several cancer types revealed that tumours from the same anatomical site can exhibit different mutation profiles, while cancers from different tissues can share driver genes and mutational profiles (Hoadley et al. 2014). Therefore, besides the mere identification of genetic variations and the discovery of cancer

driver and passenger genes, genomic characterisation of cancer has facilitated its molecular classification.

Advances in the molecular profiling of tumour tissues have helped in developing a personalised medicine approach, whereby cancer treatment is adjusted according to the molecular aberrations of individual tumours. Personalised treatment allows the selection of the most potent and effective therapy, while simultaneously sparing the patient from ineffective and costly treatments. In this context, numerous genetic changes led to the development of therapeutic agents over the past two decades. Targeting the BCR-ABL translocation with small-molecule inhibitors, such as imatinib, in chronic myeloid leukemia (CML) has been particularly successful (Deininger and Druker 2003; Druker 2003; Druker et al. 2001; Druker, Talpaz, et al. 2001). Imatinib (Glivec), the FDA approved tyrosine kinase inhibitor with activity against BCR-ABL protein, improved CML patient survival and is considered a paradigm of targeted therapies (Hochhaus et al. 2009; Druker 2004; Lydon and Druker 2004). Moreover, due to its ability to additionally inhibit the KIT protein and platelet-derived growth factor receptor, imatinib has also been used to treat gastro-intestinal tumours (Blanke et al. 2008; Verweij et al. 2004). HER-2 overexpression is also used to inform patient stratification in breast cancer and anti-HER2-specific treatment has been shown to improve survival in these patients (Piccart-Gebhart et al. 2005; Vogel et al. 2002). Similarly, KRAS mutations are used to guide therapy for colorectal carcinoma patients (Van Cutsem et al. 2011, 2009). KRAS, in contrast to other genes that are direct targets of therapies, is a molecular biomarker for anti-EGFR therapy. Patients carrying *KRAS* mutations showed no response to cetuximab, an *EGFR* inhibitor (Lièvre et al. 2006). Overall, use of molecular biomarkers has led to

tumour-specific therapeutic approaches and improved patient outcomes (Douillard et al. 2010; Mok et al. 2009).

Until recently oesophageal cancer was treated as one disease and no sub-classification of patients was implemented to guide its clinical management. In fact, many clinical trials even treated oesophageal and gastric cancer as one disease, enrolling patients with both diseases (Kopp and Hofheinz 2016; Woo, Cohen, and Grim 2015; Young and Chau 2016). However, genomic characterisation revealed that oesophageal squamous cell carcinoma resembled squamous cell carcinomas of other tissues more than OAC, whereas OAC strongly resembled chromosomally unstable gastric cancer (The Cancer Genome Atlas Research Network 2017). These data further proved that clinical management of OAC is not optimal and provided a possible explanation for the disappointing and inconclusive results of most clinical trials (Woo, Cohen, and Grim 2015; Young and Chau 2016; Kopp and Hofheinz 2016).

Efforts to characterise molecular subgroups in OAC (and gastric cancer) that would respond favourably to specific treatment led to the identification of *HER2* amplification in approximately 20% of tumours. Trastuzumab, a monoclonal antibody targeting *HER2*, was approved for clinical use based on results of the phase III clinical trial ToGA (Bang et al. 2010), which enrolled 594 previously untreated HER2-positive patients with oesophagogastric cancer. Overall, trastuzumab not only significantly improved the response rate from 35 to 47% of cases, but also prolonged the median survival of these patients from 11.1 to 13.8 months (Kopp and Hofheinz 2016), emphasising the importance of patient stratification in clinical management.

However, there are no effective therapeutic interventions for patients without HER2 amplification, and patients might acquire resistance to trastuzumab or other inhibitors of receptor tyrosine kinases. Therefore, a recent study (Secrier et al. 2016) sought to investigate the use of mutational signatures (Alexandrov et al. 2013) for stratification of OACs. Based on this, Secrier et al. defined three groups of OACs – namely DDR impaired, C>A/T dominant and mutagenic - with distinct mutational patterns and aetiologies. In the DDR-impaired group (~18% of OACs) defects in homologous recombination and chromosome segregation pathways were discovered, suggesting a possible benefit from synthetic lethality-based therapeutic approaches using PARP inhibitors. Moreover, patients in the mutagenic group would probably benefit from immunotherapy-based interventions as they exhibited higher nonsynonymous mutation and neoantigen burden than the rest of OACs (Secrier et al. 2016). Patients in the C>A/T dominant group would continue to be treated with conventional chemotherapy until more progress is made on this group of patients. Further studies will be needed for pre-clinical validation of these patient subgroups before their implementation in the clinical management of OAC.

Since OAC helper genes predicted with sysSVM perturb similar pathways (chapter 3), I sought to use these pathways to define OAC subgroups and test whether these subgroups are related to clinical features. Finally, I collaborated with two wet-lab scientists[6] to experimentally validated several helpers in each OAC subgroup to investigate the tumorigenic potential of these perturbations.

[6] Experiments were performed by Dr. Lorena Benedetti and Dr. Elizabeth Foxall

4.3 Results

4.3.1 Cancer helpers reveal six molecular subgroups of OAC patients

The discovery of cancer helpers and the fact that they converged towards perturbations of the same biological processes (chapter 3) allowed patient stratification using perturbed processes instead of individual genes. As cancer helpers are mainly rare or patient-specific, I clustered OACs according to the proportion of perturbed pathways they had in common, calculating the Jaccard index between each pair of samples. I then used hierarchical clustering to group pairs of samples with similar values of Jaccard indices (see Methods). To dissect and compare the contribution of helpers and known drivers separately, I performed two gene set enrichment analyses considering both gene sets. As described in chapter 3, in this cohort of 261 OACs, I manually annotated 202 known driver genes from the Cancer Gene Census and predicted 952 cancer helper genes using sysSVM. Overall, 212 and 189 pathways out of the 1,877 tested were found enriched for known drivers and helpers, respectively. Then, I calculated the similarity of perturbed pathways for each pair of OACs (*i* and *j*) using the Jaccard index for both known drivers (Jaccard D_{ij}) and helpers (Jaccard H_{ij}) (Figure 4.1). Finally, I used hierarchical clustering to group samples based on the values of Jaccard D and Jaccard H and find subgroups of OACs with the same number (and identity) of perturbed pathways. The results of the hierarchical clustering are reported in figure 4.2 (see below).

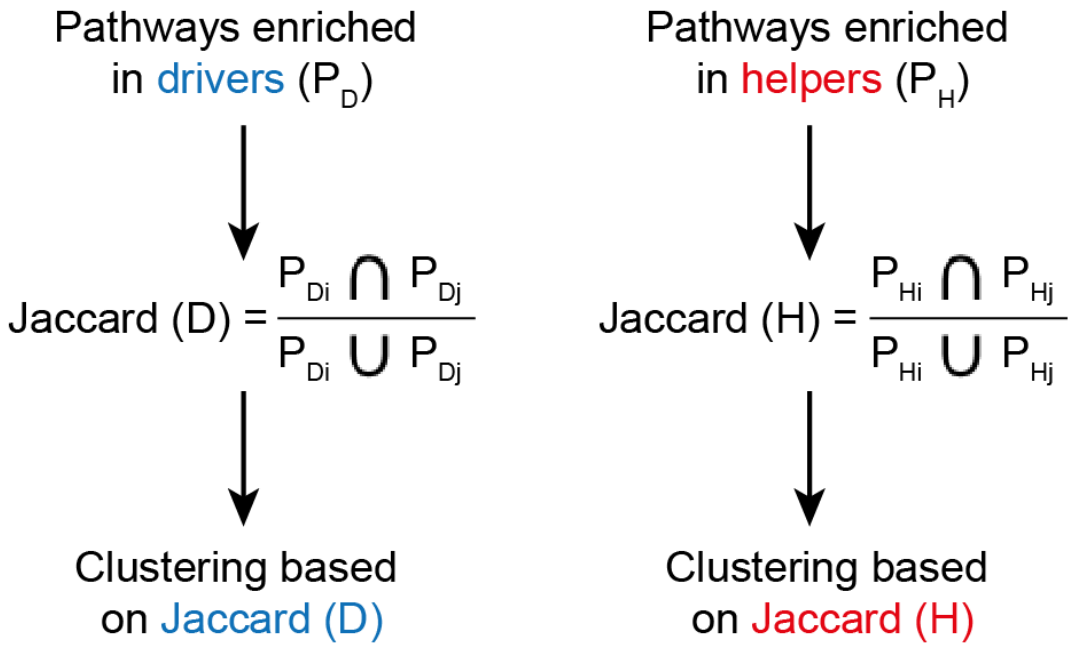


Figure 4.1. Patient stratification using shared perturbed processes in OACs. Schematic of the procedure to cluster OACs according to pathways enriched in known drivers or helpers. Enriched pathways are mapped to individual OACs and the Jaccard index is calculated as the proportion of shared pathways over the total pathways in each pair of samples (i, j). Hierarchical clustering was then performed.

When I used pathways enriched in known drivers, OACs were clustered based on the most recurrently altered genes and they were broadly divided in two major groups depending on *TP53* mutation status (Figure 4.2A left panel and 4.2B). Overall, I identified five statistically supported clusters of patients with median silhouette score of 0.5 (Figure 4.3A). These clusters were driven by the mutational status of recurrent known drivers, such as *EGFR*, *ERBB2* and *MYC in* cluster 1D, and *MYC* and *KRAS* in cluster 2D (Figure 4.2B). OACs with no alterations in *TP53* (clusters 4D and 5D) showed an overall lower mutational

burden (p = 0.03, Wilcoxon rank sum test), fewer known drivers and consequently a lower number of enriched pathways (p = $7x10^{-6}$, Wilcoxon rank sum test). Therefore, patient stratification using known drivers was confounded by the mutation status of *TP53*.

When I grouped OACs according to the pathways enriched in helpers, I identified six well-supported clusters (clusters 1H-6H, Figure 4.2A right panel) with a median silhouette score of 0.3 (Figure 4.3B). In contrast to the patient groups that were derived using known drivers, OACs were brought together not by the mutational status of *TP53* (or other recurrent known drivers), but by several helpers mapping to the same or related pathways (Table 4.1). For example, both clusters 1H and 3H showed perturbations in intracellular signalling (Figure 4.2A, Table 4.1), often involving “universal” cancer pathways (perturbed in more than 50% of OACs, see chapter 3; Figure 4.4). In a large fraction of OACs (~45%) comprising clusters 1H and 3H, the perturbations in universal cancer pathways occurred in samples with no known drivers. This suggested that more patients can be associated with perturbations in well-known cancer-related pathways if helpers were considered. Other pathways perturbed in cluster 1H, but not in 3H, involved cell cycle regulation, Toll-like receptor (TLR) signalling and proteasome activity (Figure 4.4). Finally, OACs in cluster 3H were enriched in tobacco smokers (Figure 4.5A), but no smoking mutational signature could be recovered when their mutations were analysed.

Similar to what I observed in clusters 1H and 3H, the biological pathways perturbed in clusters 2H and 4H were also functionally related. In this case, both exhibited alterations in cell cycle regulation (Figure 4.2A, Table 4.1, Figure 4.4). All OACs in cluster 2H had helpers involved in the regulation of G1/S transition, such

as members of the E2F family of transcription factors and their associated co-activators, competitors and downstream targets (Table 4.1). Cluster 4H instead harboured perturbations in DNA replication, with alterations in the MCM complex (Table 4.1), which is a downstream target of E2F (Ohtani et al. 1999; Yoshida and Inoue 2004). Dysregulation of E2F transcription factors or the MCM complex can increase genomic instability through either aberrant cell-cycle control or replicative stress (Hills and Diffley 2014; Nath et al. 2015). Consistently with this, OACs in clusters 2H and 4H exhibited high genomic instability. In particular, samples in 2H had a significantly higher number of genes with damaging somatic point mutations, indels and amplifications when compared to the rest of OACs (Figure 4.5A). Samples in 4H harboured a significantly higher number of somatically deleted genes (Figure 4.5A). The mutational profiles of OACs in cluster 2H showed significant enrichment in mutational signature 2 (attributed to the activity of AID/APOBEC family of cytidine deaminases), while OACs in cluster 4H showed enrichment in mutational signature 3 (associated with germline and somatic *BRCA1* and *BRCA2* mutations) (Figure 4.5A). Finally, OACs in cluster 4H had significantly lower survival time compared to the rest of the cohort (Figure 4.5B). Interestingly, elevated expression of the MCM complex has been previously associated with poor patient survival in multiple tumour types, including oesophageal squamous cell carcinoma, but not adenocarcinoma (Giaginis et al. 2010). The perturbation of MCM proteins and their related pathways might therefore contribute to tumour aggressiveness and poor outcome in patients of this cluster. Taken together, the results of this thesis associate MCM proteins (mainly *MCM7*) with reduced survival in oesophageal adenocarcinoma.

Cluster 5H showed perturbations in the Toll-like receptor signalling cascade (Table 4.1, Figure 4.4) that has recently been reported to be dysregulated in OAC (Fels Elliott et al. 2017). In our cohort of 261 OACs, cluster 5H accounted for 11.1% (n=29) of OACs. TLR cascades are critical for host-microbe interactions and genes in this pathway have been found mutated in other solid tumours exposed to microbial communities, such as uterine endometrial carcinoma (Husseinzadeh and Davenport 2014) and stomach adenocarcinoma (Schmaußer et al. 2005). Therefore, a better understanding of the contribution of TLR signalling cascades to tumorigenesis and the inflammation of tumour microenvironment in OAC could inform potential therapeutic interventions.

Overall, clusters 1H to 5H accounted for 166 OACs (64% of the total cohort) (Table 4.1). The remaining 95 OACs in cluster 6H shared fewer perturbed pathways than OACs in the other five groups, but, nevertheless, they exhibited shared perturbations in pathways related to Rho GTPase activity. Specifically, 55 of them (58%) had alterations in pathways related to Rho GTPase activity (Table 4.1, Figure 4.4) with modifications of Rho GTPase effectors, such as *ROCK1*, *PTK2*, *PAK1*, *LIMK1* and *NDE1*.

The purpose of hierarchical clustering using the Jaccard index for both known drivers and helpers (Figure 4.2A) was to examine whether helpers could lead to patient subgroups that were not identifiable using known drivers. Results showed that helper-derived OAC groups were indeed dispersed in the clustering analysis using known drivers (Figure 4.2A). This indicated that helpers brought together OACs with similar perturbed processes, a finding that could not have been appreciated if the analysis was merely focused on recurrent drivers. Thus, highlighting that helpers might have a value for patient stratification.

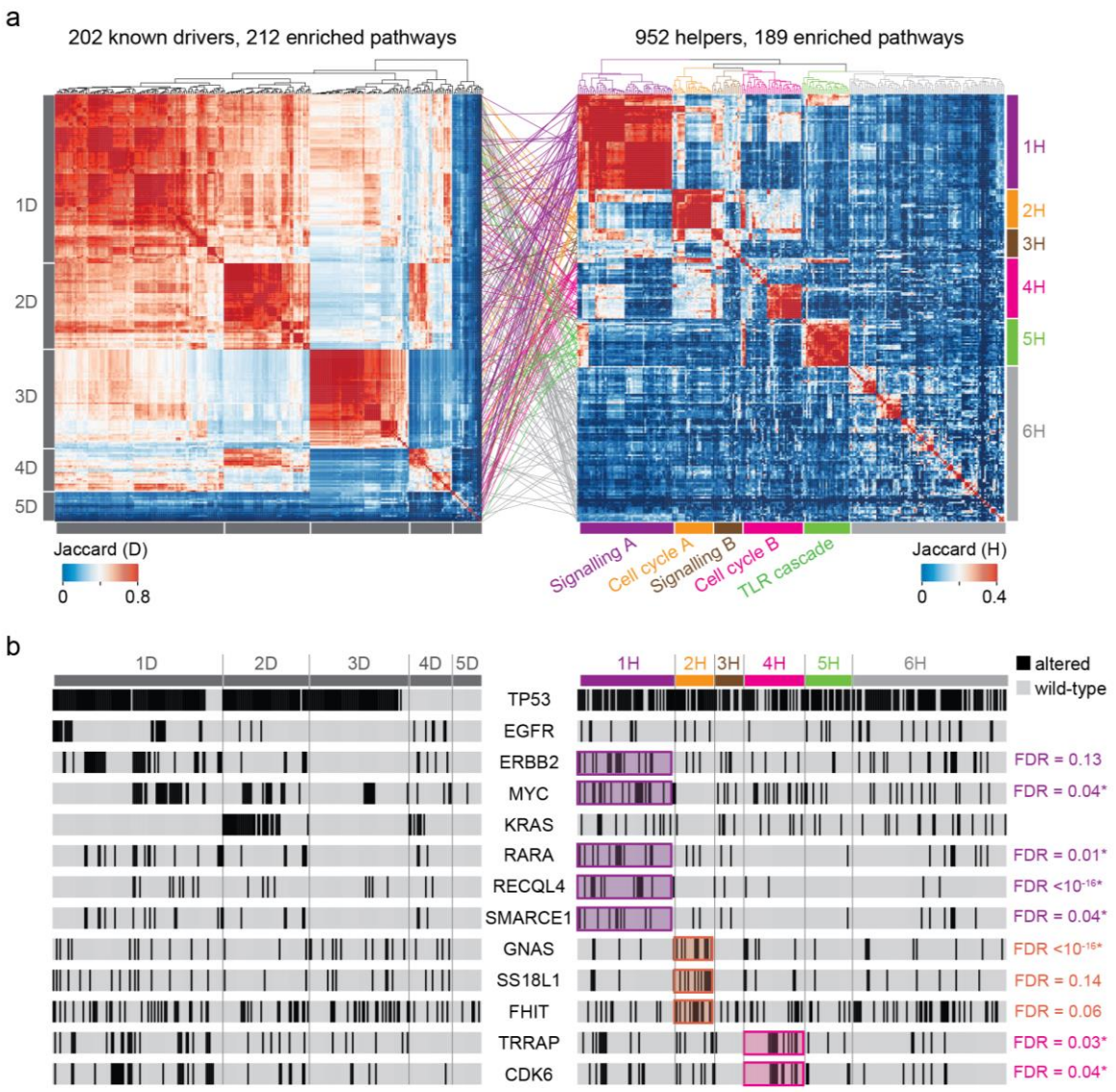


Figure 4.2. Perturbed processes in 261 OACs. **(a)** Clustering of 261 OACs according to pathways enriched in known drivers and helpers. Five clusters were identified using known drivers (1D-5D) and six using helpers (1H-6H). Cluster-matching coloured lines show where OACs clustered by pathways enriched in helpers map in the driver clusters. **(b)** Mutational status of selected known drivers across 261 OACs. Drivers enriched in clusters of helpers are highlighted and their associated significance is reported on the right-hand side of the plot. Significance was assessed using Fisher's exact test, after corrected for False Discovery Rate (FDR). Gene with FDR < 0.05 are marked with asterisk.

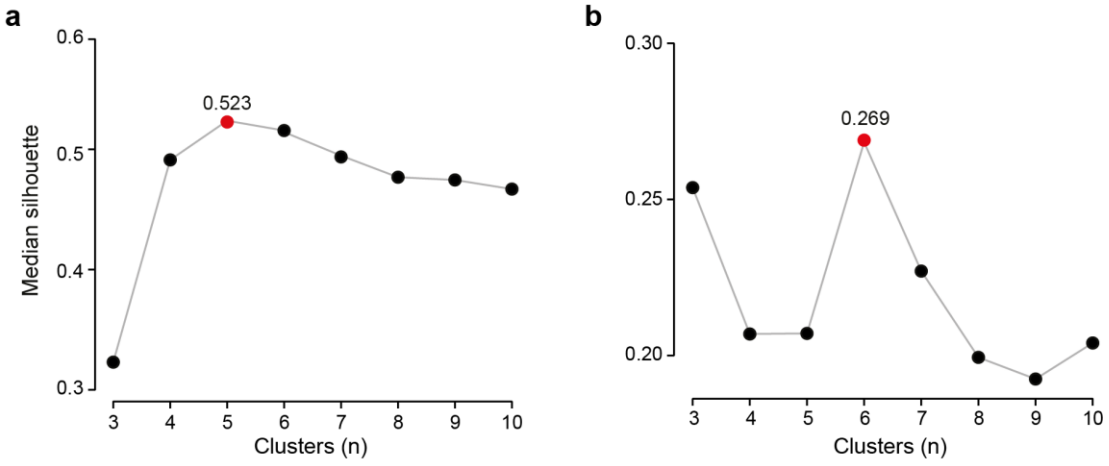


Figure 4.3. Identification of the optimal number of clusters. Silhouette analysis to measure clustering robustness of **(a)** known drivers and **(b)** helper genes. For each number of clusters between 3 and 10, clusters were derived from the dendrogram (Figure 4.2A) and the silhouette value (Rousseeuw 1987) was then calculated for each sample using the Euclidean distance between rows of the Jaccard matrix A_{ij}. The number of clusters with the highest median silhouette value over all samples was chosen as the most robust clustering partition.

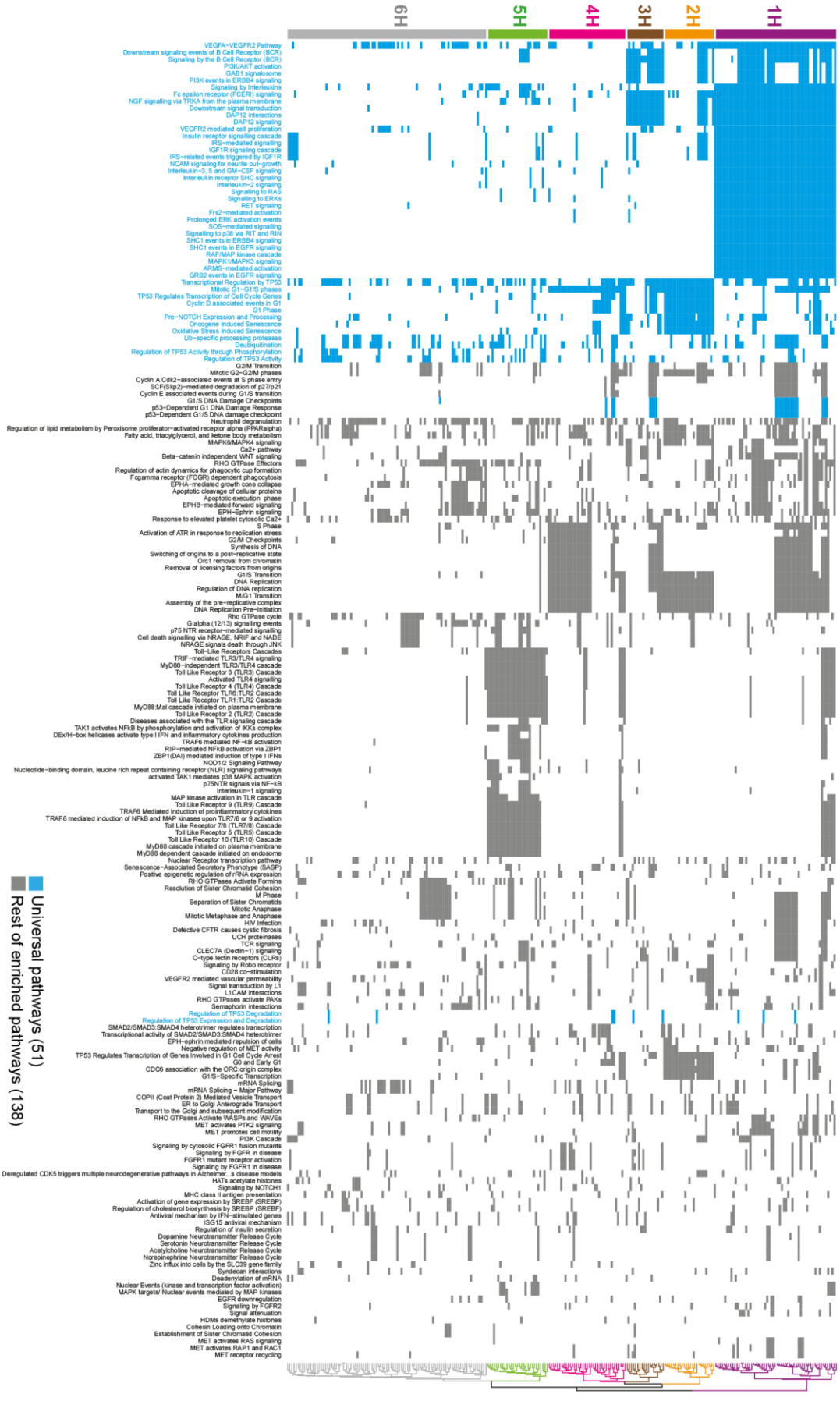

6H
5H
4H
3H
2H
1H
VEGFA-VEGFR2 Pathway
Downstream signaling events of B Cell Receptor (BCR)
Signaling by the B Cell Receptor (BCR)
PI3K/AKT activation
GAB1 signalosome
PI3K events in ERBB4 signaling
Signaling by Interleukins
Fc epsilon receptor (FCER) signaling
NGF signalling via TRKA from the plasma membrane
Downstream signal transduction
DAP12 interactions
DAP12 signaling
VEGFR2 mediated cell proliferation
Insulin receptor signalling cascade
IRS-mediated signalling
IGF1R signaling cascade
IRS-related events triggered by IGF1R
NCAM signaling for neurite out-growth
Interleukin-3, 5 and GM-CSF signaling
Interleukin receptor SHC signaling
Interleukin-2 signaling
Signalling to RAS
Signalling to ERKs
RET signaling
Frs2-mediated activation
Prolonged ERK activation events
SOS-mediated signalling
Signalling to p38 via RIT and RIN
SHC1 events in ERBB4 signaling
SHC1 events in EGFR signaling
RAF/MAP kinase cascade
MAPK1/MAPK3 signaling
ARMS-mediated activation
GRB2 events in EGFR signaling
Transcriptional Regulation by TP53
Mitotic G1-G1/S phases
TP53 Regulates Transcription of Cell Cycle Genes
Cyclin D associated events in G1
G1 Phase
Pre-NOTCH Expression and Processing
Oncogene Induced Senescence
Oxidative Stress Induced Senescence
Ub-specific processing proteases
Deubiquitination
Regulation of TP53 Activity through Phosphorylation
Regulation of TP53 Activity
G2/M Transition
Mitotic G2-G2/M phases
Cyclin A:Cdk2-associated events at S phase entry
SCF(Skp2)-mediated degradation of p27/p21
Cyclin E associated events during G1/S transition
G1/S DNA Damage Checkpoints
p53-Dependent G1 DNA Damage Response
p53-Dependent G1/S DNA damage checkpoint
Neutrophil degranulation
Regulation of lipid metabolism by Peroxisome proliferator-activated receptor alpha (PPARalpha)
Fatty acid, triacylglycerol, and ketone body metabolism
MAPK6/MAPK4 signaling
Ca2+ pathway
Beta-catenin independent WNT signaling
RHO GTPase Effectors
Regulation of actin dynamics for phagocytic cup formation
Fcgamma receptor (FCGR) dependent phagocytosis
EPHA-mediated growth cone collapse
Apoptotic cleavage of cellular proteins
Apoptotic execution phase
EPHB-mediated forward signaling
EPH-Ephrin signaling
Response to elevated platelet cytosolic Ca2+
S Phase
Activation of ATR in response to replication stress
G2/M Checkpoints
Synthesis of DNA
Switching of origins to a post-replicative state
Orc1 removal from chromatin
Removal of licensing factors from origins
G1/S Transition
DNA Replication
Regulation of DNA replication
MG1 Transition
Assembly of the pre-replicative complex
DNA Replication Pre-Initiation
Rho GTPase cycle
G alpha (12/13) signalling events
p75 NTR receptor-mediated signalling
Cell death signalling via NRAGE, NRIF and NADE
NRAGE signals death through JNK
Toll-Like Receptors Cascades
TRIF-mediated TLR3/TLR4 signaling
MyD88-independent TLR3/TLR4 cascade
Toll Like Receptor 3 (TLR3) Cascade
Activated TLR4 signalling
Toll Like Receptor 4 (TLR4) Cascade
Toll Like Receptor TLR6:TLR2 Cascade
Toll Like Receptor TLR1:TLR2 Cascade
MyD88:Mal cascade initiated on plasma membrane
Toll Like Receptor 2 (TLR2) Cascade
Diseases associated with the TLR signaling cascade
TAK1 activates NFkB by phosphorylation and activation of IKKs complex
DEx/H-box helicases activate type I IFN and inflammatory cytokines production
TRAF6 mediated NF-kB activation
RIP-mediated NFkB activation via ZBP1
ZBP1(DAI) mediated induction of type I IFNs
NOD1/2 Signaling Pathway
Nucleotide-binding domain, leucine rich repeat containing receptor (NLR) signaling pathways
activated TAK1 mediates p38 MAPK activation
p75NTR signals via NF-kB
Interleukin-1 signaling
MAP kinase activation in TLR cascade
Toll Like Receptor 9 (TLR9) Cascade
TRAF6 Mediated Induction of proinflammatory cytokines
TRAF6 mediated induction of NFkB and MAP kinases upon TLR7/8 or 9 activation
Toll Like Receptor 7/8 (TLR7/8) Cascade
Toll Like Receptor 5 (TLR5) Cascade
Toll Like Receptor 10 (TLR10) Cascade
MyD88 cascade initiated on plasma membrane
MyD88 dependent cascade initiated on endosome
Nuclear Receptor transcription pathway
Senescence-Associated Secretory Phenotype (SASP)
Positive epigenetic regulation of rRNA expression
RHO GTPases Activate Formins
Resolution of Sister Chromatid Cohesion
M Phase
Separation of Sister Chromatids
Mitotic Anaphase
Mitotic Metaphase and Anaphase
HIV Infection
Defective CFTR causes cystic fibrosis
UCH proteinases
TCR signaling
CLEC7A (Dectin-1) signaling
C-type lectin receptors (CLRs)
Signaling by Robo receptor
CD28 co-stimulation
VEGFR2 mediated vascular permeability
Signal transduction by L1
L1CAM interactions
RHO GTPases activate PAKs
Semaphorin interactions
Regulation of TP53 Degradation
Regulation of TP53 Expression and Degradation
SMAD2/SMAD3:SMAD4 heterotrimer regulates transcription
Transcriptional activity of SMAD2/SMAD3:SMAD4 heterotrimer
EPH-ephrin mediated repulsion of cells
Negative regulation of MET activity
TP53 Regulates Transcription of Genes Involved in G1 Cell Cycle Arrest
G0 and Early G1
CDC6 association with the ORC:origin complex
G1/S-Specific Transcription
mRNA Splicing
mRNA Splicing - Major Pathway
COPII (Coat Protein 2) Mediated Vesicle Transport
ER to Golgi Anterograde Transport
Transport to the Golgi and subsequent modification
RHO GTPases Activate WASPs and WAVEs
MET activates PTK2 signaling
MET promotes cell motility
PI3K Cascade
Signaling by cytosolic FGFR1 fusion mutants
Signaling by FGFR in disease
FGFR1 mutant receptor activation
Signaling by FGFR1 in disease
Deregulated CDK5 triggers multiple neurodegenerative pathways in Alzheimer...s disease models
HATs acetylate histones
Signaling by NOTCH1
MHC class II antigen presentation
Activation of gene expression by SREBF (SREBP)
Regulation of cholesterol biosynthesis by SREBP (SREBF)
Antiviral mechanism by IFN-stimulated genes
ISG15 antiviral mechanism
Regulation of insulin secretion
Dopamine Neurotransmitter Release Cycle
Serotonin Neurotransmitter Release Cycle
Acetylcholine Neurotransmitter Release Cycle
Norepinephrine Neurotransmitter Release Cycle
Zinc influx into cells by the SLC39 gene family
Syndecan interactions
Deadenylation of mRNA
Nuclear Events (kinase and transcription factor activation)
MAPK targets/ Nuclear events mediated by MAP kinases
EGFR downregulation
Signaling by FGFR2
Signal attenuation
HDMs demethylate histones
Cohesin Loading onto Chromatin
Establishment of Sister Chromatid Cohesion
MET activates RAS signaling
MET activates RAP1 and RAC1
MET receptor recycling
Universal pathways (51)
Rest of enriched pathways (138)

Figure 4.4. OAC clustering using pathways enriched in helpers. Hierarchical clustering was performed using a presence/absence matrix of perturbed pathways and OACs as described in Methods. The order of OACs along the y axis corresponds to that shown in figure 4.2A for helpers, including the six clusters (1H-6H). Samples were assigned to a given pathway if they had at least one altered helper mapping to that pathway. Fifty-one of the 73 universal pathways perturbed in at least 50% of OACs are coloured in light blue. All other colours depict cluster-defining pathways**.**

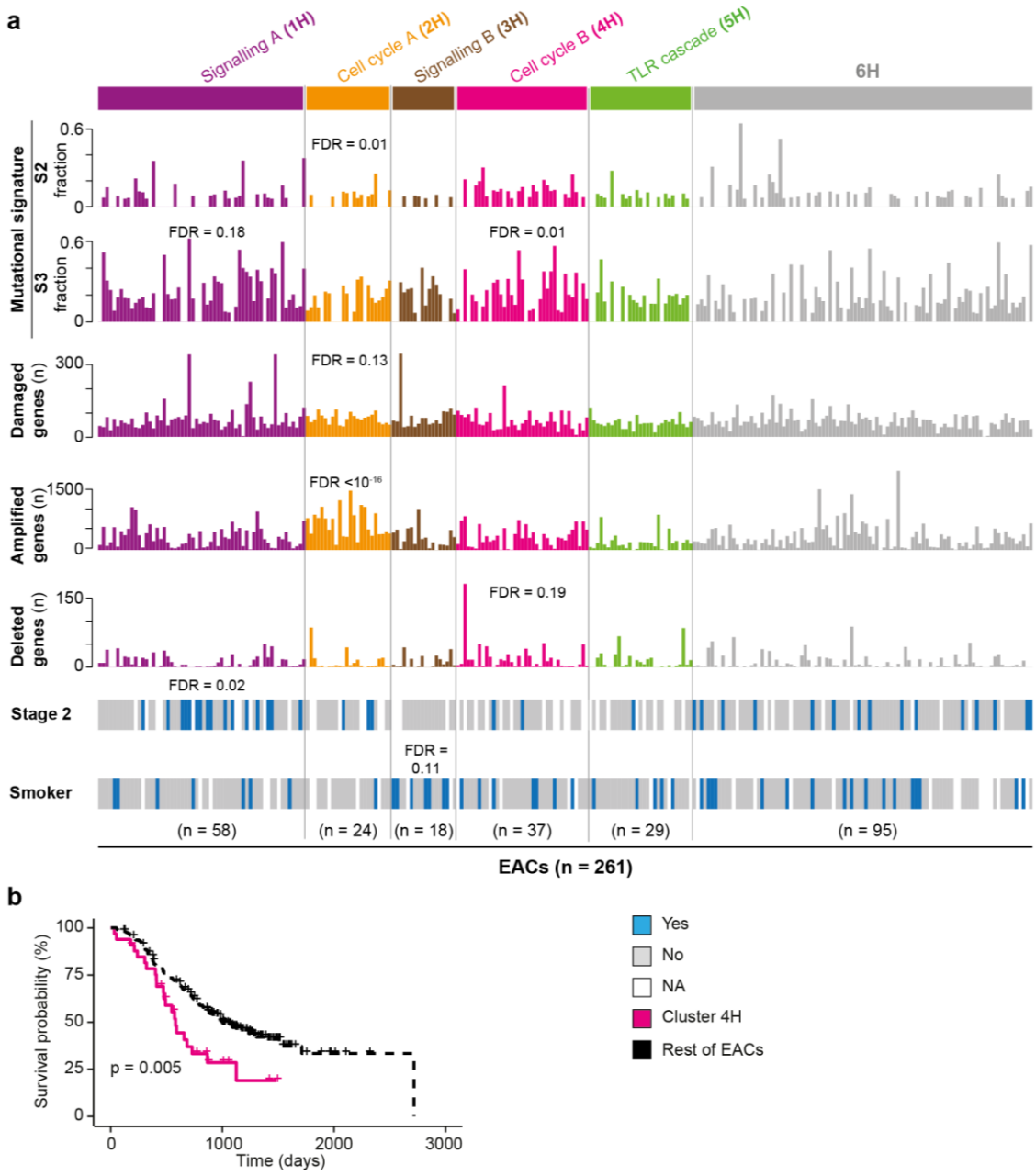


Figure 4.5. Features of OAC clusters driven by pathways enriched in helpers. For each helper cluster (1H-6H) indicated are the molecular features (mutational signatures, number of genes with damaging mutations, undergoing amplification or deletion), the distribution of stage 2 tumours and the tobacco

smoking habits of the patients that show significant associations with one of the six clusters of helpers. Enrichment in number of altered genes, tumour staging and smoking habits was assessed using Fisher's exact test. Distributions of mutational signatures were compared using Wilcoxon rank-sum test. FDR = false discovery rate after correction for multiple testing. **b.** Kaplan-Meier survival curves of OACs in cluster 4H (n = 37) and the rest of OACs (n = 224). Analysis was performed using survival and survminer R packages with default parameters. Significance was measured using the log-rank test.

Table 4.1. Summary of patient clusters derived using cancer helpers. For each cluster the number of samples, a set of representative genes and pathways are reported. For genes and pathways, the number in parenthesis denotes the corresponding number of OACs with perturbations. The full data set of genes and pathways is reported in Appendix Tables 7.2 and 7.3.

Cluster	Samples in cluster	Representative helpers	Representative pathways
1H	58	PTK2(14),SCRIB(14),AGO2(13), THRA(11),ASAP1(9),MCM7(7), MED24(7),SPTBN1(7),STK3(6), BAG6(5)	ARMS-mediated activation(58),DAP12 interactions(58),DAP12 signalling(58), Downstream signal transduction(58),Fc epsilon receptor (FCERI) signalling(58),Frs2-mediated activation(58),GRB2 events in EGFR signalling(58),IGF1R signalling cascade(58),Insulin receptor signalling cascade(58),Interleukin receptor SHC signalling(58)
2H	24	E2F1(23),TOMM34(20),DNMT3B(17), NCOA3(17),VAPB(10),PTPN1(7), STK4(6), DC25B(5),MMP9(5),RBL1(5)	Assembly of the pre-replicative complex(24),CDC6 association with the ORC:origin complex(24),Cyclin D associated events in G1(24),DNA Replication(24),DNA Replication Pre-Initiation(24),G1 Phase(24),M/G1 Transition(24),Mitotic G1-G1/S phases(24),Oncogene Induced Senescence(24),Oxidative Stress Induced Senescence(24)
3H	18	BAG6(5),PRRC2A(5),CDKN1A(4), ASAP1(3),ATP7B(3),EHMT2(3), ITPR1(3),AGO1(2),AGO2(2),CTPS1(2)	DAP12 interactions(18),DAP12 signalling(18),Downstream signal transduction(18),Fc epsilon receptor (FCERI) signalling(18),NGF signalling via TRKA from the plasma membrane(18),Signalling by the B Cell Receptor (BCR)(18),Downstream signalling events of B Cell Receptor (BCR)(15),GAB1 signalosome(15),PI3K events in ERBB4 signalling(15),PI3K/AKT activation(15)
4H	37	MCM7(18),PLOD3(9),CUX1(6), NCOA3(6),TOMM34(6),VAPB(5), ASAP1(4),MMP9(4),RBL1(4), SLC25A13(4)	Mitotic G1-G1/S phases(37),G1/S Transition(29),DNA Replication(26),Regulation of DNA replication(26),S Phase(26),Assembly of the pre-replicative complex(24),DNA Replication Pre-Initiation(24),G2/M Checkpoints(24),M/G1 Transition(24),Orc1 removal from chromatin(23)
5H	29	MEF2C(5),HMGB1(4),HSPH1(4), IKBKB(4),IRAK1(4),NBEA(4),APP(3), CABLES1(3),DYRK2(3),FLNA(3)	Activated TLR4 signalling(29),MyD88:Mal cascade initiated on plasma membrane(29),Toll Like Receptor 2 (TLR2) Cascade(29),Toll Like Receptor 4 (TLR4) Cascade(29),Toll Like Receptor TLR1:TLR2 Cascade(29),Toll Like Receptor TLR6:TLR2 Cascade(29),Toll-Like Receptors Cascades(29),MyD88-independent TLR3/TLR4 cascade(28),Toll Like Receptor 3 (TLR3) Cascade(28),TRIF-mediated TLR3/TLR4 signalling(28)
6H	95	DLC1(14),PAK4(12),POGZ(11), LONRF1(10),ROCK1(10),VEGFA(10), CERS2(9),ACTN4(8),DYRK1B(8), ABCC1(7)	Neutrophil degranulation(46),RHO GTPase Effectors(38),VEGFA-VEGFR2 Pathway(36),Transcriptional Regulation by TP53(35),EPH-Ephrin signalling(33),Rho GTPase cycle(31),Deubiquitination(29),Fatty acid, triacylglycerol, and ketone body metabolism(28),G alpha (12/13) signalling events(28),Regulation of TP53 Activity(27)

4.3.2 Helper-defined OAC subgroups are associated with specific perturbations of known drivers

The initial hypothesis of this work was that sysSVM predictions could have a helper role in OAC. Consequently, I hypothesised that helpers might occur concomitantly with specific known drivers. To understand the dynamics of helpers and known drivers, I searched for over-represented altered known drivers in each one of the OAC subgroups described above.

OACs in cluster 1H were significantly associated with five known drivers (*RECQL4*, *RARA*, *MYC*, *SMARCE1* and *ERBB2*; Figure 4.2B), which were often, but not always, co-amplified. Interestingly, *SCRIB*, a recurrent helper in cluster 1H (Table 4.1) was recently found to inhibit liver cancer cell proliferation by suppressing the expression of *MYC* (Kapil et al. 2017), suggesting a functional association between a predicted helper and a known driver. However, although this finding suggested a tumour-suppressing role, in OAC, *SCRIB* was predicted as an amplified helper (Appendix Table 7.2), an indication of a tumour-promoting role. Cluster 2H was characterised by significant alterations of the known drivers *GNAS*, *SS18L1*, and *FHIT* (Figure 4.2B). *FHIT* is linked to increased genomic instability (Saldivar et al. 2012) and regulates the expression of cell cycle-related genes (Weiske, Albring, and Huber 2007), therefore potentially affecting the G1/S transition pathways of this cluster (Table 4.1). Finally, cluster 4H showed frequent alterations in the known drivers *TRAPP* and *CDK6* (Figure 4.2B). The latter functions in various cell cycle-related pathways, including the mitotic G1/S phase pathway that was found altered in 100% of cluster 4H (Table 4.1, Figure 4.4).

Taken together, these results showed that pathways perturbed by helpers were associated with alterations of specific known drivers in at least three of the six OAC clusters that were identified. This dynamic relationship of known drivers and helpers, although hypothesised in the initial stages of this thesis, is well supported by my data. Further investigation is warranted to dissect the exact interaction of helpers and these known drivers across different OAC subgroups.

Finally, to test whether patient stratification was affected by considering only the top 10 helper genes in each OAC, I performed the same analysis considering as helpers the top five or top 15 scoring genes (528 and 1,297 unique genes, respectively). By altering the arbitrary cut-off of top 10 scoring genes, I confirmed that both pathway enrichment analysis and the corresponding OAC subgroups were reproducible across multiple sets of high-scoring genes from sysSVM. In particular, the vast majority of pathways enriched in top five and top 15 (99% and 77%) were also enriched in the top 10 scoring genes (Figure 4.6A, B). This indicated that the recurrently perturbed processes were highly overlapping among different sets of top scoring genes. I then clustered OACs according to the proportion of shared perturbed pathways, as described before in Figure 4.1. Thereby, I verified that the six clusters obtained using pathways enriched in top 10 genes recapitulated well the clusters obtained using pathways enriched in top five or top 15 genes (Figure 4.6C, D). Therefore, the clustering was robust regardless of the applied ranking cut-off.

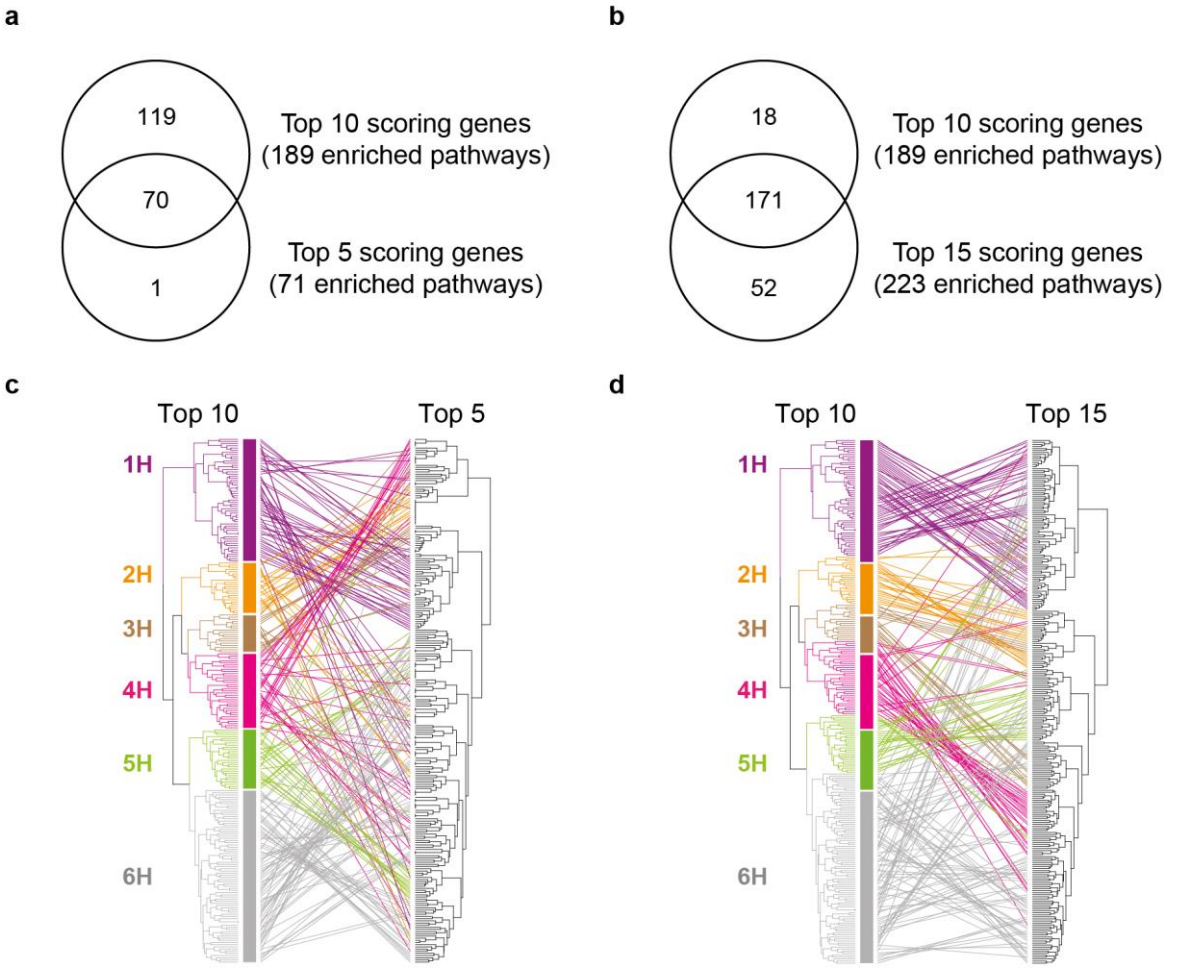


Figure 4.6. Comparison of helpers using different ranking cut offs. Comparison of enriched pathways between top ten and (**a**) top five or (**b**) top 15 scoring genes in each sample. Gene set enrichment analysis using top five and top 15 scoring genes led to 71 and 223 enriched pathways, respectively (FDR <0.01). Comparison of sample positions in the clustering dendrograms of top 10 and (**c**) top five or (**d**) top 15 scoring genes in each sample. Complete linkage hierarchical clustering with Euclidean distance was used to group 261 OACs according to pathways enriched in the different datasets of helpers. The dendrogram of top 10 scoring genes corresponds to that shown in Figure 4.2A.

4.3.3 Pan-cancer prevalence of alterations in cell-cycle-related helpers

The dysregulation of cell-cycle is a hallmark of cancer cells. In this cohort of 261 OACs, sysSVM predicted several helpers that were part of cell cycle regulation pathways. In particular, transcription factors of the *E2F* gene family and members of MCM complex were found recurrently altered in OAC clusters 2H and 4H (Table 4.1). Therefore, they were selected for experimental validation (see below) and pan-cancer analysis in order to dissect the prevalence and mode of their alteration in multiple cancer types.

E2F family of transcription factors consists of eight proteins (E2F1-8), which can be further sub-classified in activators (E2F1-3) and repressors (E2F4-8) based on their downstream effect on gene transcription (Ogawa et al. 2002). Activator E2Fs promote cell proliferation, while repressors control cell cycle exit and differentiation (Ishida et al. 2001). Amplification and overexpression of E2Fs (mainly *E2F1*) has been detected in multiple cancer types, such as ductal breast cancer and non-small cell lung cancer (Montenegro et al. 2014; Lu et al. 2012; Hung et al. 2012). Although often discovered overexpressed, E2F1 is thought to play a dual role in cancer cells. There is accumulating evidence that apart from its oncogenic role via overexpression, it also acts as a tumour suppressor by inducing apoptotic cell death upon perturbation of normal cell cycle control (Johnson 2000).

In OAC, sysSVM predicted *E2F1* as an amplified helper in 23 out of 24 samples of cluster 2H (9% of the total cohort; Table 4.1). Pan-cancer analysis of 7,828 TCGA patients revealed that *E2Fs* were amplified in various frequencies across different cancer types. Overall, more than 50% of the samples had at least one *E2F* amplified in 15 out of 31 cancer types examined in TCGA (Figure 4.7A).

Although this frequency might be slightly overestimated, as the copy number of genes was not corrected with the ploidy of each sample (see Methods), it clearly demonstrated that *E2F*s were frequently amplified in multiple cancer types. The highest frequency was observed in uveal melanoma, rectum adenocarcinoma, uterine carcinosarcoma, ovarian serous cystadenocarcinoma and adrenocortical carcinoma, while acute myeloid leukaemia, kidney renal clear cell carcinoma, thymoma and thyroid carcinoma had a low fraction of samples with *E2F* amplifications (Figure 4.7A). When the prevalence of amplification of each *E2F* was examined separately, *E2F5* was the most frequently amplified gene (53% of patients), while *E2F2* was the least amplified gene (5% of patients) (Figure 4.7B). Overall, from a total of 3,284 patients with at least one *E2*F amplified, 1,961 (60%) patients harboured amplification in only one *E2F* gene, suggesting a possible mutual exclusivity of amplification events in multiple members of the *E2F* gene family.

From the inspection of altered *E2F*s across different cancer types, interesting patterns of *E2F* amplifications were revealed (Figure 4.7C). Both *E2F1* and *E2F5* were found frequently amplified in colon and rectum cancers, kidney chromophobe, ovarian serous carcinoma, uterine carcinosarcoma, uveal melanoma and adrenocortical carcinoma. Frequent amplifications of *E2F5*, but not *E2F1* were observed in head and neck squamous cell carcinoma, breast invasive carcinoma and testicular germ cell tumours. *E2F4*, although rarely amplified in the pan-cancer cohort (8%; Figure 4.7B), was frequently amplified in adrenocortical carcinoma, kidney chromophobe and kidney renal papillary cell carcinoma. On the other hand, very low number of samples were found with *E2*F amplifications in brain tumours (lower grade glioma and glioblastoma), kidney renal clear cell carcinoma, prostate

and pancreatic adenocarcinomas. Consistent with my results in the OAC cohort, oesophageal carcinoma, showed high number of patients with *E2F1* amplifications (Figure 4.7C). Taken together, these observations suggest that there are four groups of cancer types with respect to amplifications of *E2*Fs: i) those with amplifications of both *E2F*1 and *E2F3*, ii) those with only *E2F5* amplifications, ii) those with only *E2F1* amplifications and iv) those with very low frequency of amplifications of *E2F*s.

Another cycle-related amplified helper that was predicted by sysSVM in 18 out of the 37 OACs (7% of the total cohort; Table 4.1) in cluster 4H was *MCM7*. MCM proteins (MCM2-7) are part of the minichromosome maintenance complex, which serves as the eukaryotic replicative helicase that unwinds double-stranded DNA and promotes the formation of DNA replication forks (Bochman and Schwacha 2009). *MCM* genes have been found in both eukaryotes and archaea and they share significant sequence similarity mainly around the 250-amino-acid region that encodes the ATPase active site (Koonin 1993). The MCM complex is loaded to the origins of DNA replication during the G_1 phase of the cell cycle and it is subsequently activated by cyclin-dependent kinases (CDKs) and Dbf4-dependant kinase (DDK) to promote the assembly of the replication forks (Nougarède et al. 2000). Inactivation of any of the MCMs blocks DNA elongation and leads to the loss of integrity of the replication fork. There is substantial evidence that fine-tuning of phosphorylation and dephosphorylation of MCMs regulate cell cycle and protects genome integrity, while aberrant phosphorylation associates with uncontrolled proliferation and the development of multiple cancers (Fei and Xu 2018). Recent studies that evaluated the expression of *MCM*s across various cancers types reported their overexpression in non-small cell lung cancer

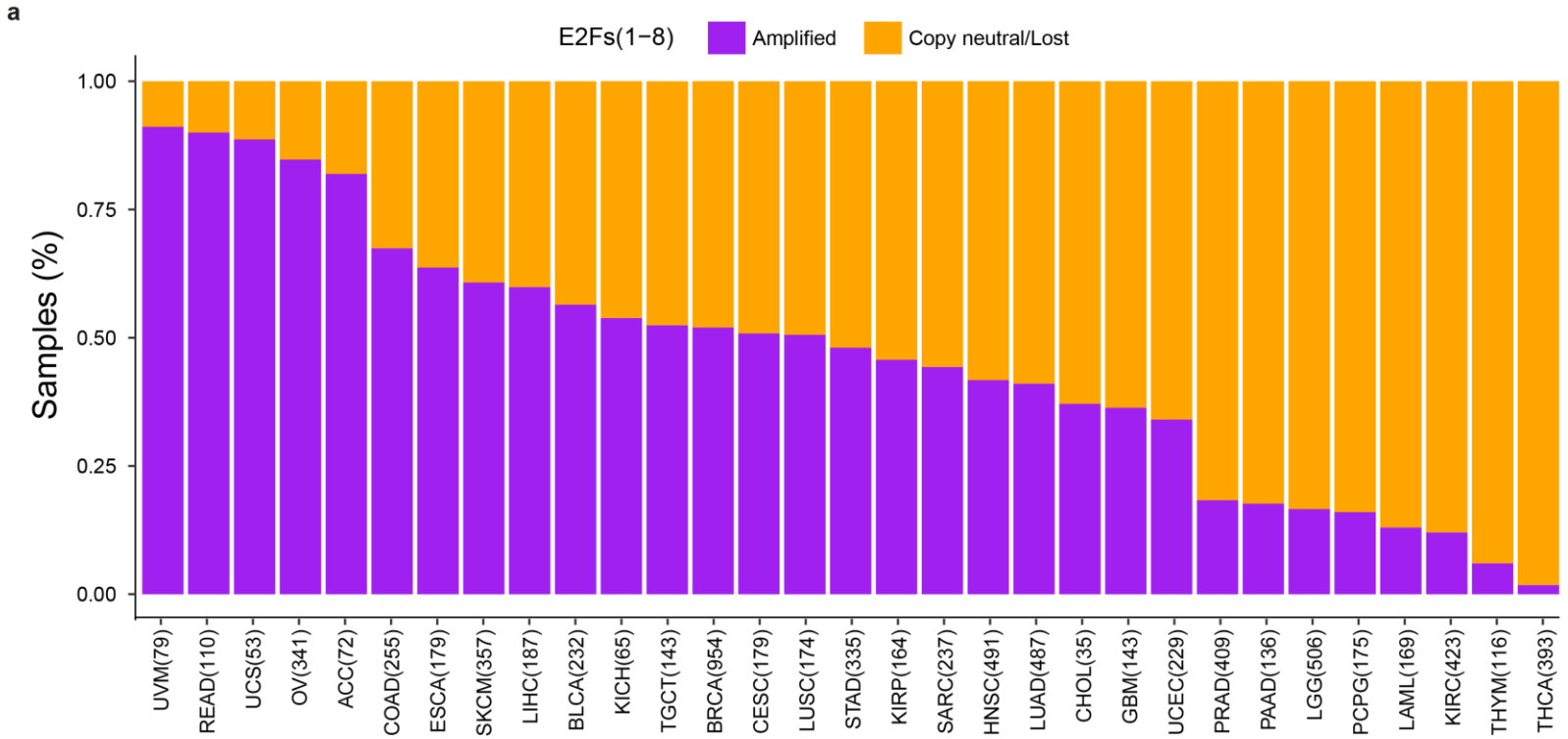
(Liu et al. 2017) and breast cancer (Kwok et al. 2015) and their association with shorter overall survival time.

*MCM*s were found frequently amplified in many cancer types in the pan-cancer cohort (Figure 4.8A). Some of the cancer types with high frequency of *MCM* amplifications, such as uterine carcinosarcoma, ovarian serous carcinoma and adrenocortical carcinoma, also harboured frequent *E2F* amplifications (Figure 4.7A). Interestingly, other cancer types, such as glioblastoma and lung squamous cell carcinoma, that harboured no *E2F* amplifications, showed quite frequent amplifications of *MCM*s (Figure 4.8A). This suggested that aberrations in these two regulators of the cell cycle (i.e. *E2F*s and *MCM*s) were not always mutually exclusive, as observed in OAC and glioblastoma, but they could also co-occur.

Amplification of *MC*Ms was associated with overexpression in the pan-cancer cohort and the level of expression of different members of the MCM complex was co-regulated, as patients with at least one component amplified tended to have all components overexpressed (Figure 4.8B). This suggested that *MCM* amplification had a functional effect in cancer cells, which led to the elevated expression of the MCM complex as a whole and eventually increased proliferation rate (see next paragraph).

Taken together, these results showed that amplification of the components of cell cycle is a frequent event across multiple cancer types. This, in combination with results from recent literature, further supports the fact that these predictions of sysSVM are involved in tumorigenesis. In OAC, amplifications of *E2F*s and *MCM*s were predicted as helpers in more than 15% of the total cohort and these amplifications were linked to overexpression (see next paragraph). In the remaining part of this chapter, I describe the experimental validation of these

genes, and others, in an effort to investigate the magnitude of their tumorigenic effect in multiple oesophageal cell lines.


a
E2Fs(1−8)
Amplified
Copy neutral/Lost
Samples (%)
1.00
0.75
0.50
0.25
0.00
UVM(79)
READ(110)
UCS(53)
OV(341)
ACC(72)
COAD(255)
ESCA(179)
SKCM(357)
LIHC(187)
BLCA(232)
KICH(65)
TGCT(143)
BRCA(954)
CESC(179)
LUSC(174)
STAD(335)
KIRP(164)
SARC(237)
HNSC(491)
LUAD(487)
CHOL(35)
GBM(143)
UCEC(229)
PRAD(409)
PAAD(136)
LGG(506)
PCPG(175)
LAML(169)
KIRC(423)
THYM(116)
THCA(393)

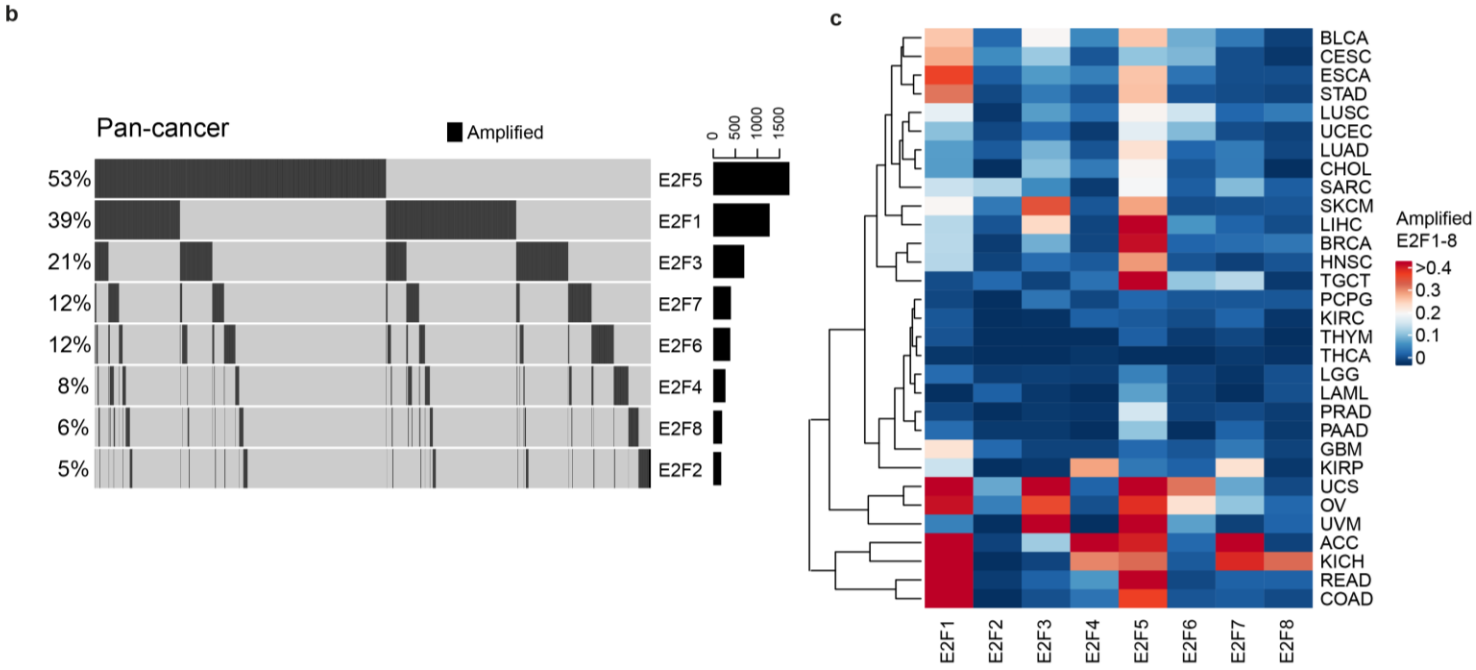
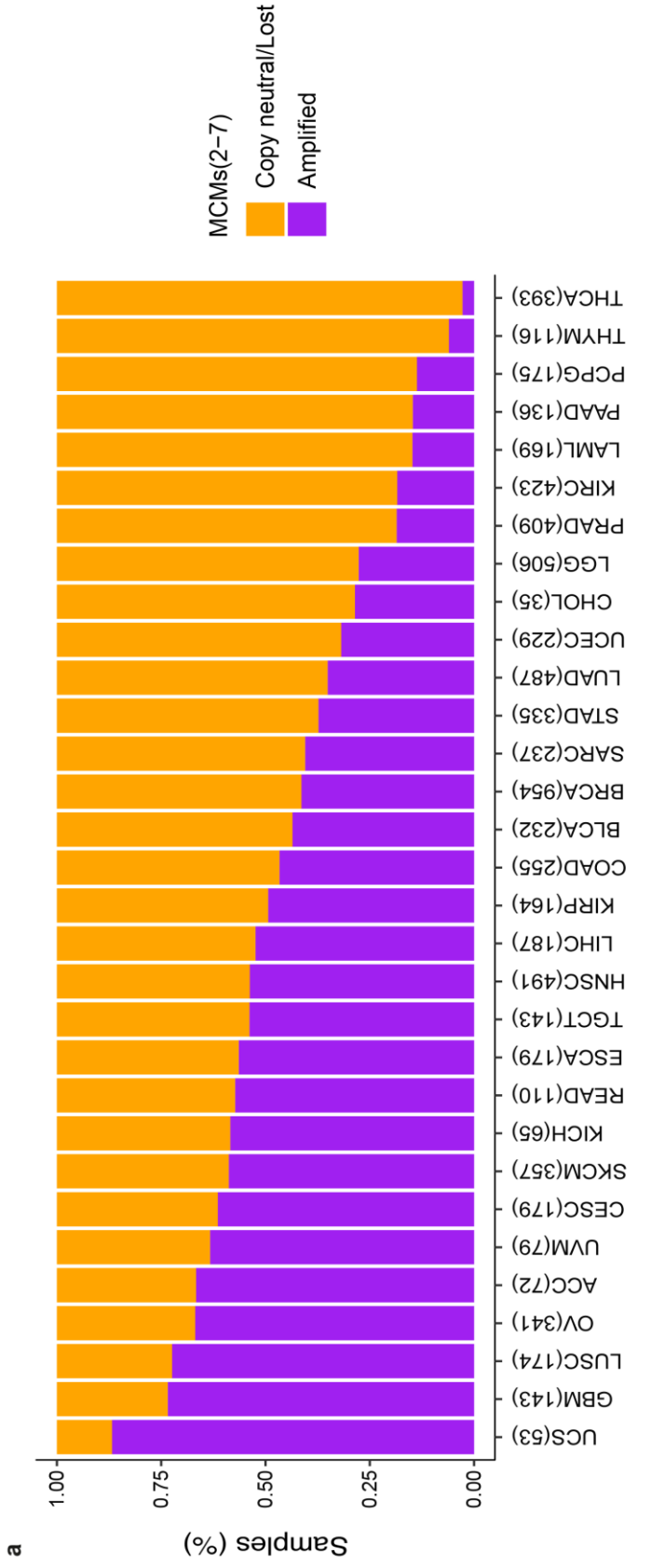


Figure 4.7. Pan-cancer analysis of amplification events associated with E2F transcription factors. **(a)** Percentage of patients with amplified E2Fs (E2F1-8) across 31 cancer types from TCGA. The cohort was comprised of 7,828 patients and amplifications were defined as described in Methods. **(b)** Pan-cancer prevalence of E2F amplifications for all 8 genes of the E2F family of transcription factors. **(c)** Pan-cancer prevalence of E2F amplifications across different cancer types from TCGA. Cancer type abbreviations are summarised in table 3.5.


a
Samples (%)
1.00
0.75
0.50
0.25
0.00
MCMs(2−7)
Copy neutral/Lost
Amplified
UCS(53)
GBM(143)
LUSC(174)
OV(341)
ACC(72)
UVM(79)
CESC(179)
SKCM(357)
KICH(65)
READ(110)
ESCA(179)
TGCT(143)
HNSC(491)
LIHC(187)
KIRP(164)
COAD(255)
BLCA(232)
BRCA(954)
SARC(237)
STAD(335)
LUAD(487)
UCEC(229)
CHOL(35)
LGG(506)
PRAD(409)
KIRC(423)
LAML(169)
PAAD(136)
PCPG(175)
THYM(116)
THCA(393)

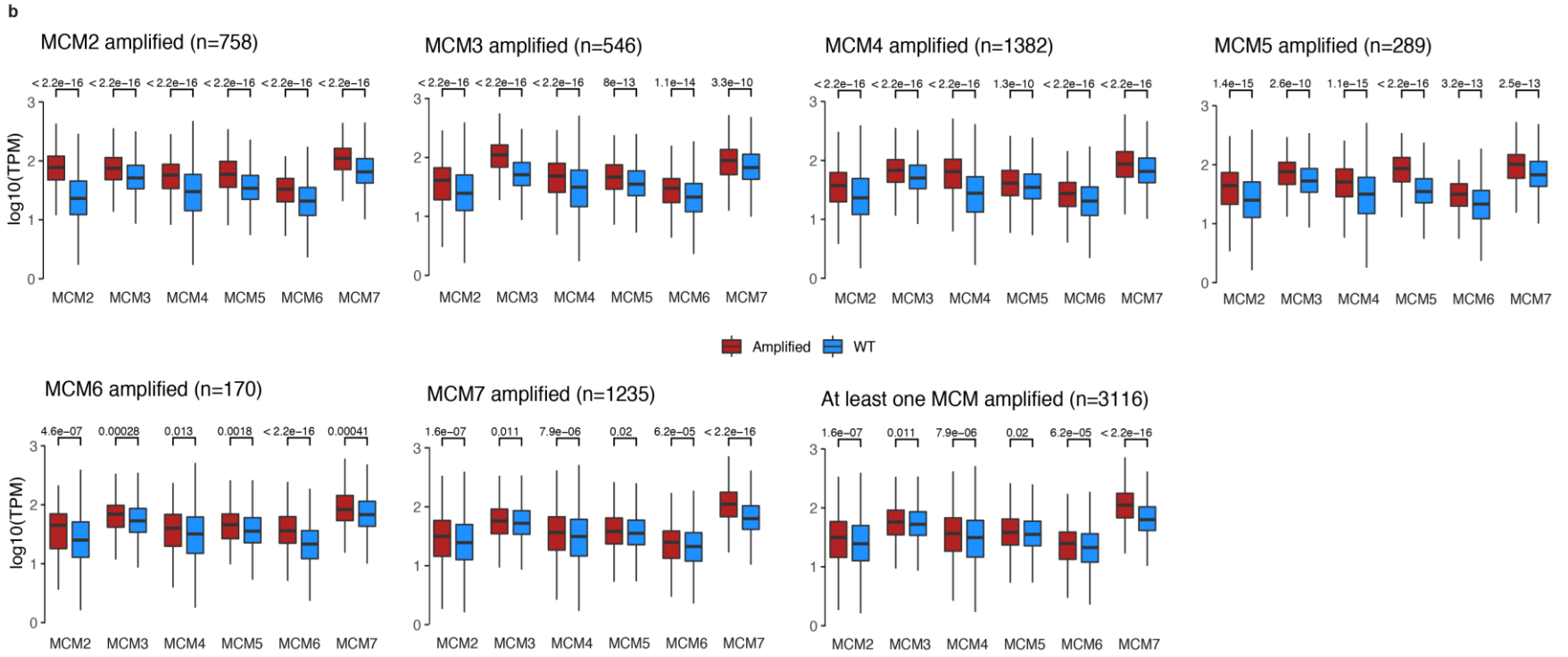


Figure 4.8. Pan-cancer analysis of amplification events associated with members of the MCM complex. **(a)** Percentage of patients with amplified MCMs (MCM2-7) across 31 cancer types from TCGA. The cohort was comprised of 7,828 patients and amplifications were defined as described in Methods. **(b)** Pan-cancer analysis of expression of MCMs in patients reported in panel a. For each MCM subunit, patients with amplification events were annotated. Then the expression of all MCM subunits was compared in those patients versus the rest of the cohort. Statistical significance was calculated using a two-tailed Wilcoxon rank sum test. Cancer type abbreviations are summarised in table 3.5.

4.3.4 Experimental validation of the commonly altered helpers *E2F1* and *MCM7*

To experimentally validate the predictions of sysSVM and investigate their role in the tumorigenesis of OAC, I collaborated with Dr. Lorena Benedetti and Dr. Elizabeth Foxall, who assessed the consequences of silencing and overexpression of representative helpers in cancer cell lines *in vitro*. As representative, we selected helpers that met all the following criteria:

1. They were frequently altered helpers in specific OAC clusters or rare helpers that contributed to perturbations of enriched pathways, across multiple clusters of OACs (as those described in figure 4.2A)
2. They were expressed in the oesophageal cancer cell lines that would be used in the experiments
3. They were not amplified in the oesophageal cancer cell lines that would be used in the experiments (optimisation of CRISPR design)

The selection of rare, as well as frequent helpers, was performed in order to confirm our working hypothesis that tumorigenic pathways aberrations can be the result of alterations of multiple genes, regardless of their frequency.

We used two experimental approaches. In the first one, we assessed the consequences of altering representative helpers in FLO-1 cells. This is an OAC diploid cell line with no mutations or copy number alterations in any of the helpers selected for validation. Therefore, it enabled a direct evaluation of the effect of the introduced gene alterations without interference from already acquired alterations in the same gene. We measured cell proliferation as a main hallmark of cancer and also performed gene-specific assays. In the second approach, we sought to investigate the dependency of cancer cells upon

perturbation of helpers. In this case, we used OAC cell lines with alterations similar to those observed in patients (see Methods).

We started by modifying the most commonly altered helpers in clusters 2H and 4H, namely *E2F1* (23 out of 24 samples in cluster 2H; Table 4.1) and *MCM7* (18 out of 37 samples in cluster 4H; Table 4.1). As mentioned in the previous paragraph, both *E2F1* and *MCM7* were amplified in OACs and their amplification led to significant overexpression, when compared to copy number neutral state (median two-fold increase, $p = 6x10^{-3}$ and $p=8x10^{-3}$, respectively Wilcoxon rank-sum test; Figure 4.9A). We therefore stably overexpressed *E2F1* and *MCM7* in FLO-1 cells to levels comparable to those observed in patients (Figure 4.9B). In both cases we observed significantly increased proliferation of overexpressing cells as compared to control cells ($p=2x10^{-4}$ and $p=9x10^{-4}$, respectively, two-tailed t-test; Figure 4.9C), validating the functional impact of the overexpression of these helpers to cell proliferation. Since *E2F1* promotes cell cycle progression, we also assessed DNA replication rate by measuring EdU (5-ethynyl-2'-deoxyuridine) incorporation to newly synthesized DNA during the cell cycle. We observed increased EdU intensity throughout S phase in *E2F1* overexpressing cells as compared to control cells ($p<10^{-4}$, Mann Whiney U test; Figure 4.10A). This suggests that E2F1 may help cancer growth by promoting S phase entry. Similar to *E2F1*, to assess the functional consequence of *MCM7* overexpression, we measured the loading of the MCM complex onto chromatin. As described in the previous paragraph, MCM complex (MCM2-7) is loaded onto chromatin to promote DNA unwinding and formation of the DNA replication fork. We observed that *MCM7* overexpressing cells displayed a lower MCM fluorescence intensity overall as compared to control cells when staining the chromatin-bound fraction for either MCM7 or

MCM3 ($p<10^{-4}$, Mann-Whitney U test; Figure 4.10B, 4.10C). This suggested that less MCM complex was loaded onto chromatin by the start of S phase. Therefore, *MCM7* overexpression leads to both increased proliferation and perturbation of MCM complex activity. These two findings seem to be conflicting as one would expect higher amount of MCM complex being loaded onto the chromatin when increased proliferation is observed. More experiments are needed to dissect the mechanism behind this observation. However, one possible explanation could be that there is a delay between the overexpression of one member of MCMs and the physiological adjustment of a cell to produce higher amount of the other MCM members and eventually assemble higher amounts of the MCM complex as a whole as expected by the stoichiometry balance that was observed in figure 4.8B.

Finally, we utilised a patient-derived cell line, MFD-1 (Garcia et al., 2016), in which *MCM7* was overexpressed (four-fold higher expression) when compared to FLO-1 (Figure 4.11A), to assess the dependency of cancer cells to *MCM7* perturbation. To reduce the expression of *MCM7* in MFD-1 cells to levels comparable to the expression of FLO-1 cells, we used doxycycline-inducible shRNA lentiviral vector (Table 4.2; Figure 4.11B). This led to a significant decrease in cell proliferation ($p = 2x10^{-5}$, two-tailed t-test; Figure 4.11C), indicating that the proliferation rate of MFD-1 cells is dependent of the *MCM7* perturbation and overexpression.

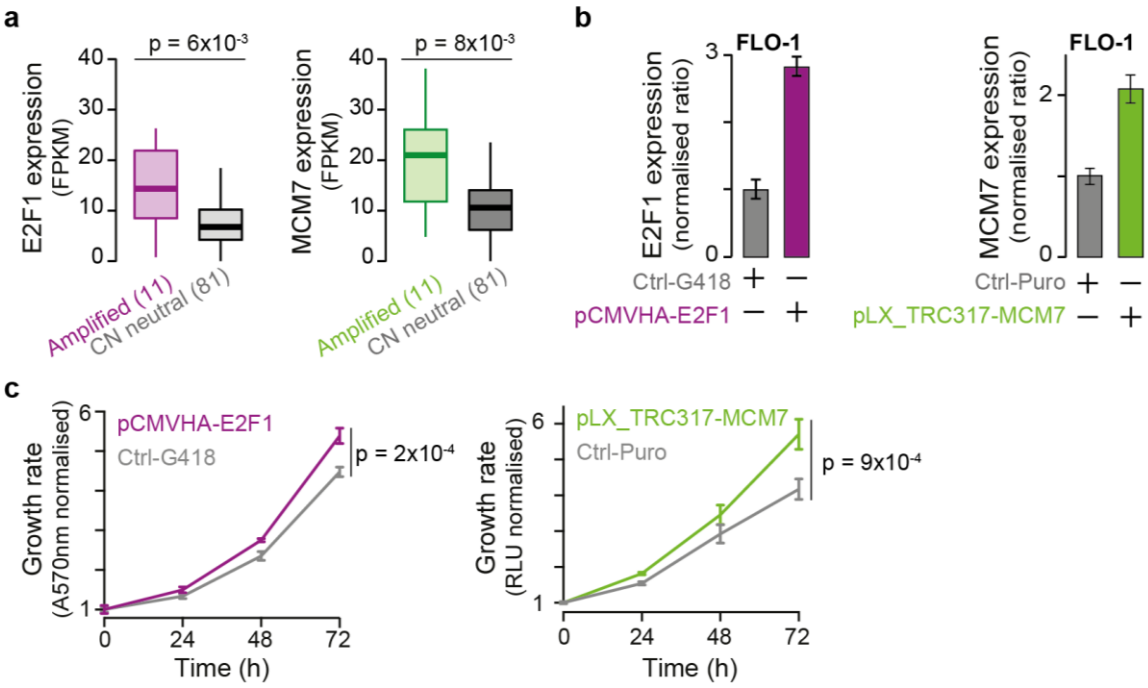


Figure 4.9. Cancer helper role of E2F1 and MCM7. **(a)** E2F1 and MCM7 expression in OACs where they are amplified (11 samples each) as compared to OACs where they are copy number neutral (81 samples each). Significance was assessed using the Wilcoxon rank-sum test. **(b)** E2F1 and MCM7 mRNA expression in FLO-1 cells assessed by qRT-PCR. Expression was relativised to β-2-microglobulin and normalised to control cells. **(c)** Proliferation curve of FLO-1 cells overexpressing E2F1 or MCM7 as compared to the corresponding control cells. Two biological replicates were performed, with reactions performed in triplicate in all qRT-PCR experiments. In proliferation assays, at least two biological replicates were performed, each with four technical replicates. Proliferation was assessed every 24 hours and each time point was normalised to time zero. Mean values at 72 hours were compared by two-tailed Student's t-test.

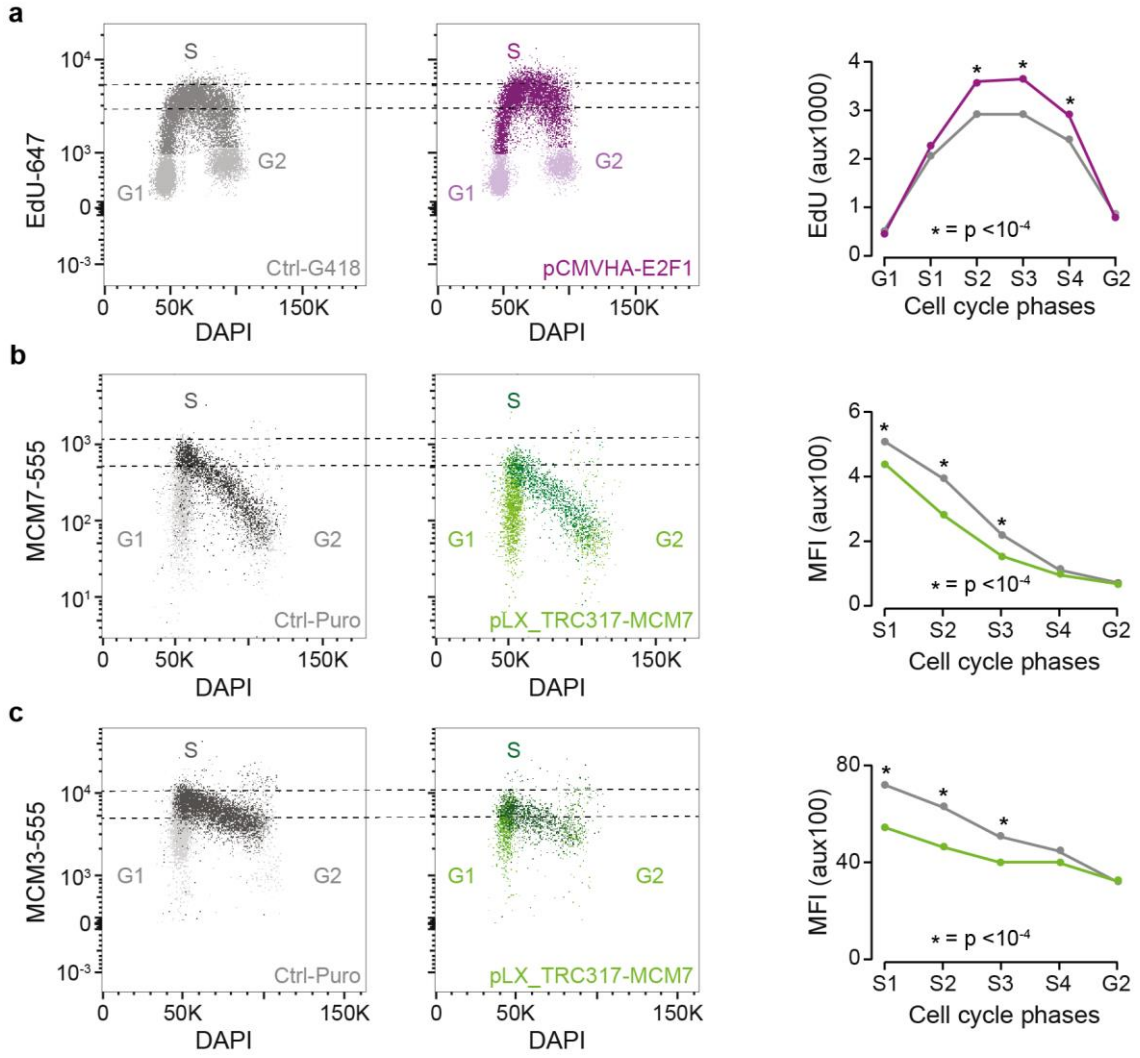


Figure 4.10. Estimation of MCM loading onto the genome during cell cycle. **(a)** Assessment of EdU (5-ethynyl-2'-deoxyuridine) incorporation by flow cytometry in E2F1 overexpressing cells as compared to control cells. Cells were separated into G1, S and G2 phases, and S phase cells were subdivided into 4 gates from early to late S phase (S1-S4). The geometric mean fluorescence intensity of EdU was measured for the cells in each gate and differences between EdU intensity were assessed using the Mann-Whitney U test. Three biological replicates were performed, and a representative experiment is shown. Quantification of MCM complex loading onto chromatin in MCM7 overexpressing or control cells via staining of MCM7 (**e**) or MCM3 (**f**). Cells were pulsed with EdU, and chromatin fractionation was performed before staining for MCM7 or MCM3 to detect the MCM complex bound to chromatin. Cells were separated into cell cycle phases using EdU and DAPI intensity (see Methods). MCM7 or MCM3 fluorescence intensity during S phase illustrates the unloading of the MCM complex from chromatin. The geometric mean fluorescence intensity of MCM staining was measured for the cells in each cell cycle gate and differences in MCM intensity were assessed using Mann-Whitney U test. Representative data from one of three biological replicates are shown.

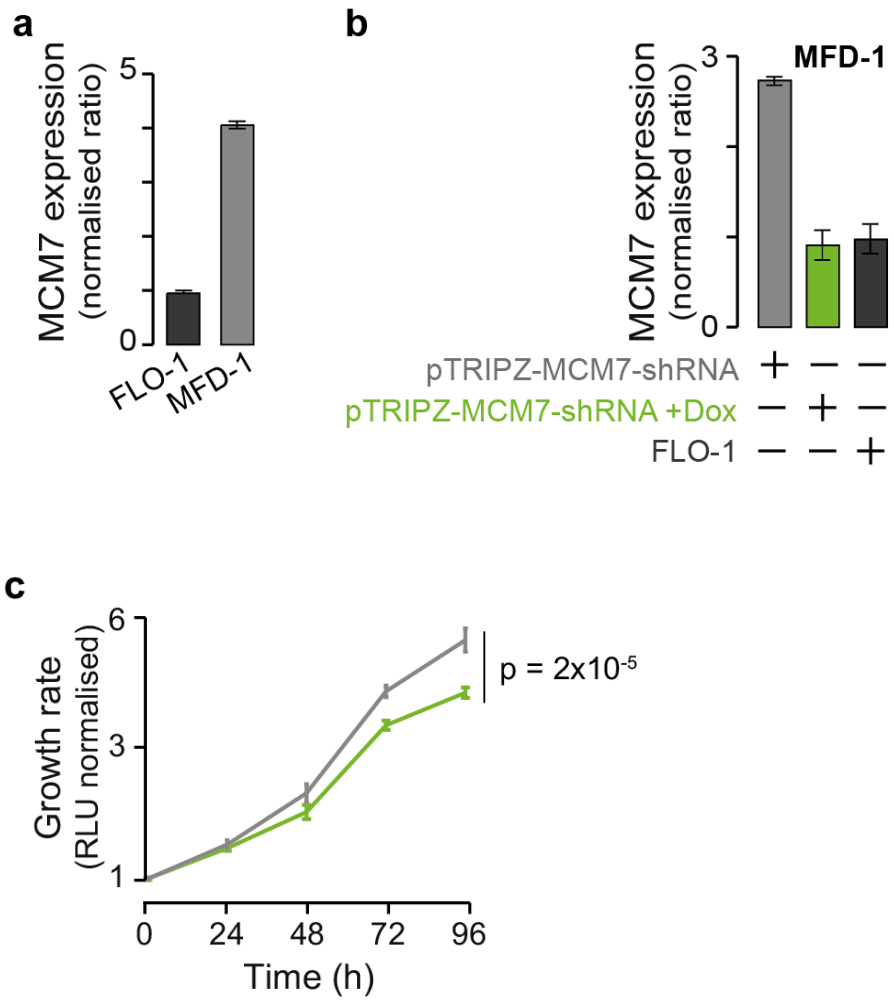


Figure 4.11. Assessment of dependency of oesophageal cancer cell lines on MCM7. **(a)** MCM7 mRNA expression levels in MFD-1 and FLO-1 cells. Expression was relativized to β-2-microglobulin and normalised to FLO-1 cells. **(b)** MCM7 expression levels in MFD-1 cells after transduction with a lentiviral vector carrying an inducible shRNA against MCM7. Expression was assessed in the absence of doxycycline and after 96 hours of doxycycline treatment, relativised to β-2-microglobulin and normalised to FLO-1 cells. **(c)** Proliferation curve of MFD-1 cells with or without doxycycline-induced MCM7 knockdown. Two biological replicates were performed, with reactions performed in triplicate in all qRT-PCR experiments. In proliferation assays, at least two biological replicates were performed, each with four technical replicates. Proliferation was assessed every 24 hours and each time point was normalised to time zero. Mean values at 72 hours were compared by two-tailed Student's t-test.

Table 4.2. List of oligos used in the study. Reported are the DNA and RNA sequences of the oligos used in this study. * = selected for knockdown experiment. NA = not applicable.

Experiment	Gene/Protein	Oligo	Sequence	Protein position (aa)
CRISPR gene editing	ABI2 (Protein ID: Q9NYB9)	ABI2_crRNA1	GGCAACACTTGCTAAGGAT	S57-A62
		ABI2_crRNA2	GCCTATCTGATAAACACCT	A62-T67
		ABI2_crRNA3	AGATTCCATCCTTCGTAGC	Q82-S88
	NCOR2 (Protein ID: Q9Y618)	NCOR2_crRNA1	TCGCTGCGGGCGGCCGACA	L361-H370
		NCOR2_crRNA2	ACCCGCTCAATGGCTAATG	R581-A590
		NCOR2_crRNA3	ACAGCGCCATCACATACCG	S1227-G1235
	NTC	NTC_crRNA1	GATACGTCGGTACCGGACCG	NA
		NTC_crRNA2	GTAACGCGAACTACGCGGGT	NA
		NTC_crRNA3	GTCGACGTTATTGCCGGTCG	NA
		NTC_crRNA4	GGAAACCTACGTCGACGAAT	NA
		NTC_crRNA5	GCTCTCGTACGGCGCGTATC	NA
MiSeq	NCOR2	NCOR2_forward1	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCCCTCGACGTAAACCACCC	NA
		NCOR2_reverse1	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCCACACTTCTCCTCTGGGG	NA
		NCOR2_forward2	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCAGTAGGTAGCGCTGGGATT	NA
		NCOR2_reverse2	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGAAGACAGACGACACCTCAGG	NA
		NCOR2_forward3	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGAGGGGTTATAAGATGGGCTGG	NA
		NCOR2_reverse3	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGACTCCCTCTGCGTTGAAAC	NA
	ABI2	ABI2_forward	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTGGACTCAGCAGAATCGTTG	NA
		ABI2_reverse	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTGCCAGCATTACAGATAGCCT	NA
Quantitative	MCM7	MCM7_forward	ATCGGATTGTGAAGATGAAC	NA

RT-PCR		MCM7_reverse	CTTTTCGTAGAAATCCTCCTC	NA
	E2F1	E2F1_forward	CTGATGAATATCTGTACTACGC	NA
		E2F1_reverse	CTTTGATCACCATAACCATCTG	NA
	PAK1	PAK1_forward	AGCTAACTGACTTTGGATTC	NA
		PAK1_reverse	GGGTTTTCATTGAGGTATGG	NA
	PSMD3	PSMD3_forward	CCTATTTCCTTCTGACTCAAG	NA
		PSMD3_reverse	CCGGATAATTAGGGTGTAGG	NA
	Cas9	Cas9_forward	GGGGGACAGTCTTCACGAGC	NA
		Cas9_reverse	CACGTACATGTCCCTGCCGT	NA
shRNA (mature antisense)	pTRIPZ-tRFP MCM7	MCM7_shRNA1, V2THS_252457	AGGTTTCTGAGAGTAAACC	NA
		MCM7_shRNA2, V3THS_340299	TTGACATCTCCATTAGCCT	NA
		MCM7_shRNA3, V3THS_340301 *	ACTTGAACTGCTTCTTCCC	NA
	pSMART-tGFP PSMD3	PSMD3_shRNA1, V3IHSMCG_4787375	CTGTCAAGGCCATGAGGTT	NA
		PSMD3_shRNA2, V3IHSMCG_5205155	GTCTCAAACAGACGGTGCA	NA
		PSMD3_shRNA3, V3IHSMCG_6080216 *	TGGGATGTCTTGGCTTGTA	NA

4.3.5 Experimental validation of rare helpers

Next, we experimentally evaluated the role of rare helpers in tumorigenesis. To this aim, we selected *NCOR2* that was altered in eight OACs across five of the six clusters (Appendix Table 7.2). In contrast to *E2F1* and *MCM7*, which were informative for patient stratification, *NCOR2* was predicted by sysSVM but, nevertheless, patients with *NCOR2* alterations were distributed across different patient clusters. *NCOR2* is part of the nuclear receptor corepressor complex that favours global chromatin deacetylation and transcriptional repression (Figure 4.12A). In accordance to previous reports that suggested a tumour suppressor role of *NCOR2* in lymphoma and prostate cancer, the most frequent *NCOR2* alterations in OAC led to loss of function. Therefore, we edited *NCOR2* in FLO-1 cells using a vector-free CRISPR system that was previously developed in our lab. Briefly, three pooled crRNAs were co-transfected with Cas9 and the tracrRNA (Methods, Table 4.2) and the editing was confirmed and quantified using Miseq (Methods, Figure 4.12B). Editing of *NCOR2* via CRISPR led to a 1.3-fold increase in proliferation rate in the edited cells compared to the control cells ($p = 3x10^{-3}$, two-tailed t-test test; Figure 4.12C), suggesting that loss-of-function perturbation of this gene confers a proliferative advantage to cancer cells.

Next, we focused on members of the Rho GTPase effector pathway, which were pervasively perturbed in all six clusters, often through patient-specific alterations (Figure 4.13A). Aberrant Rho signalling has been long linked to oncogenesis (Porter, Papaioannou, and Malliri 2016) and the perturbation of optimal signalling levels has been attributed to a wide range of mechanisms. Evidence suggests that GTPase signalling can be disrupted both directly, by Rac1 P29S mutation which was described as a driver in melanoma (Hodis et al.

2012), and indirectly via overexpression of guanine exchange factors (GEFs) (Cook, Rossman, and Der 2014) or loss of negative regulators (Wolf et al. 2003). Most of the helpers predicted by sysSVM are located downstream of Rho GTPase signalling pathway, which is supportive of their helper role. As representative genes of the Rho GTPase effectors, we modified *ABI2* and *PAK1*, which underwent damaging alterations and amplification in one and nine OACs, respectively (Appendix Table 7.2). We therefore edited *ABI2* and overexpressed *PAK1* as described above and confirmed the editing efficacy and overexpression (Figure 4.13B). The proliferation rate of cells harbouring editing of *ABI2* or overexpression of *PAK1* was assessed as before. In both cases we observed significantly increased proliferation as compared to control cells (ABI2: $p = 4x10^{-4}$, PAK1: $p = 1x10^{-3}$ two-tailed t-test; Figure 4.13C).

Finally, we focussed on *PSMD3* that encodes a subunit of the regulatory 19S proteasome complex. *PSMD3* is amplified and overexpressed in three OACs of cluster 1H, which overall contains 14 samples with alterations in six proteasome subunits (Figure 4.14A). We identified three OAC cell lines (MFD-1, OE19 and OE33) showing higher basal expression of *PSMD3* compared to FLO-1 (2-, 3- and 4-fold increase respectively, Figure 4.14B). Using a doxycycline-inducible lentiviral shRNA vector (Table 4.2), we reduced *PSMD3* expression in MFD-1, OE19 and OE33 cells to levels equivalent to those of FLO-1 (Figure 4.14C). In all three cell lines we observed a significant reduction in cell proliferation following the reduction of *PSMD3* expression (MFD-1: $p = 4x10^{-8}$; OE19: $p = 2x10^{-8}$; OE33: $p = 6x10^{-3}$, two-tailed t-test; Figure 4.14D). The effect was particularly strong in OE19, where the reduction of *PSMD3* expression to diploid levels arrested cell growth completely. In MFD-1 and OE33 it led to 1.3- and 1.2-fold reduction of cell growth (Figure 4.14D). This

suggests that the extent of OAC reliance upon helper alterations is at least partially context dependent.

Taken together, our experimental data indicated that, independently of the alteration frequency, the modification of helpers positively affected OAC cell growth. Moreover, we provided evidence that OAC cells became addicted to helper alterations, suggesting that targeting helpers, or the pathways in which they act, could reduce OAC progression.

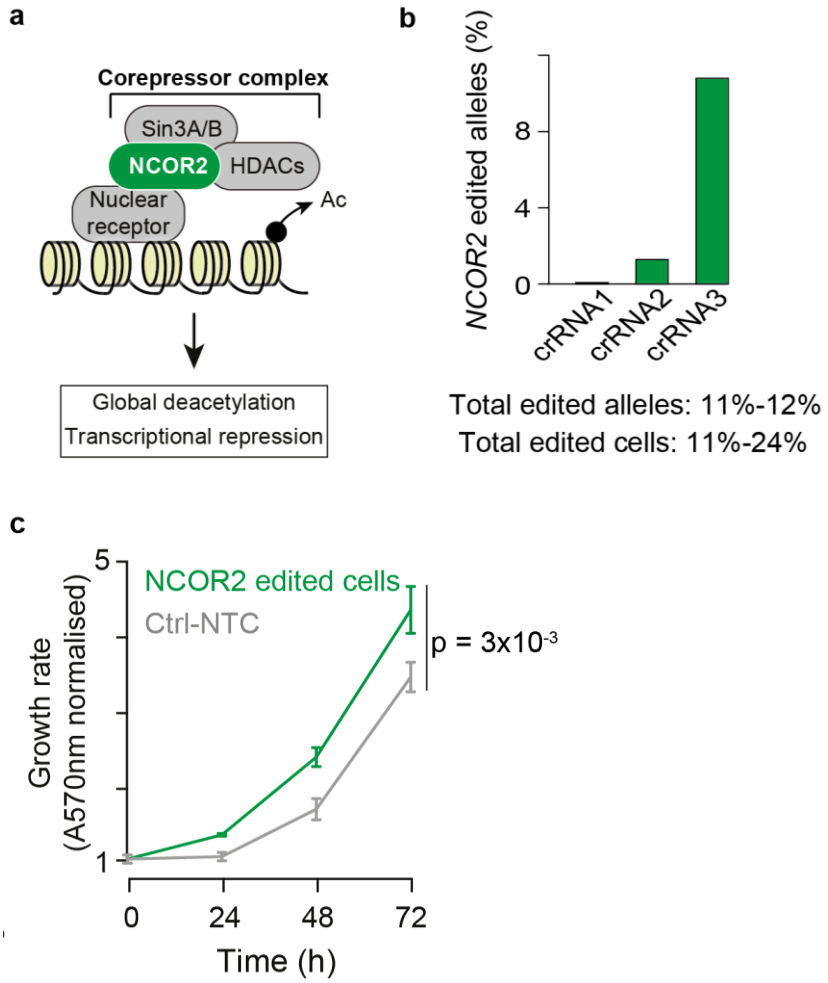


Figure 4.12. Cancer helper role of *NCOR2*. **(a)** Function of NCOR2 as part of the nuclear receptor co-repressor complex, whose activity results in chromatin deacetylation and transcriptional repression. **(b)** Editing of the *NCOR2* gene using three pooled crRNAs where cells are transiently co-transfected with Cas9 protein, crRNAs and tracrRNA (Benedetti et al. 2017). The editing efficiency was measured using Miseq and the range of edited alleles and cells was derived considering the two opposite scenarios where all three crRNAs edit the same alleles/cells or different alleles/cells, respectively. **(c)** Proliferation curve of *NCOR2* or NTC edited FLO-1 cells. Proliferation was assessed every 24 hours and each time point was normalized to time zero. Mean values at 72 hours were compared by two-tailed Student's t-test. Three biological replicates were performed, each with four technical replicates.

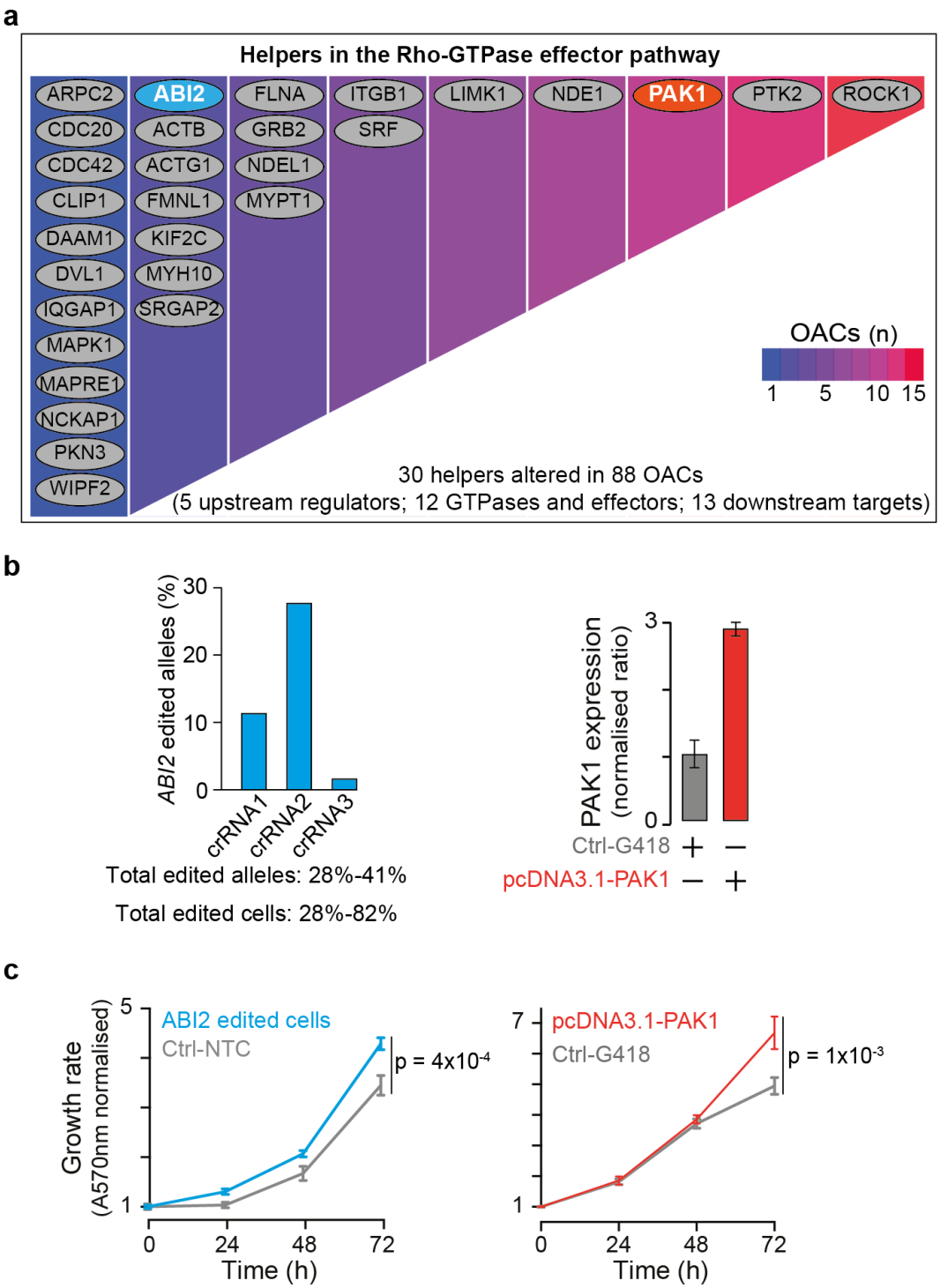


Figure 4.13. Cancer helper role of *ABI2*, and *PAK1*. **(a)** Manual curation of the helpers contributing to the Rho-GTPase effectors pathway. Heatmap indicates the number of samples with alterations in each gene. *ABI2* (blue) and *PAK1* (red) were selected for experimental validation. **(b)** Editing of the ABI2 gene (left panel) using three pooled crRNAs where cells are transiently co-transfected with Cas9 protein, crRNAs and tracrRNA (Benedetti et al. 2017). The editing efficiency was measured using Miseq and the range of edited alleles and cells was derived considering the two opposite scenarios where all three crRNAs edit the same alleles/cells or different alleles/cells, respectively. PAK1 mRNA expression in FLO-1 cells was assessed by qRT-PCR (right panel), relativized

to β-2-microglobulin and normalised to control cells. Experiments were done in triplicate in two biological replicates. **(c)** Proliferation curves of FLO-1 cells after *ABI2* editing (left panel) or *PAK1* overexpression (right panel). Three biological replicates were performed, each with four technical replicates. Proliferation was assessed every 24 hours and each time point was normalised to time zero. Mean values at 72 hours were compared by two-tailed Student's t-test.

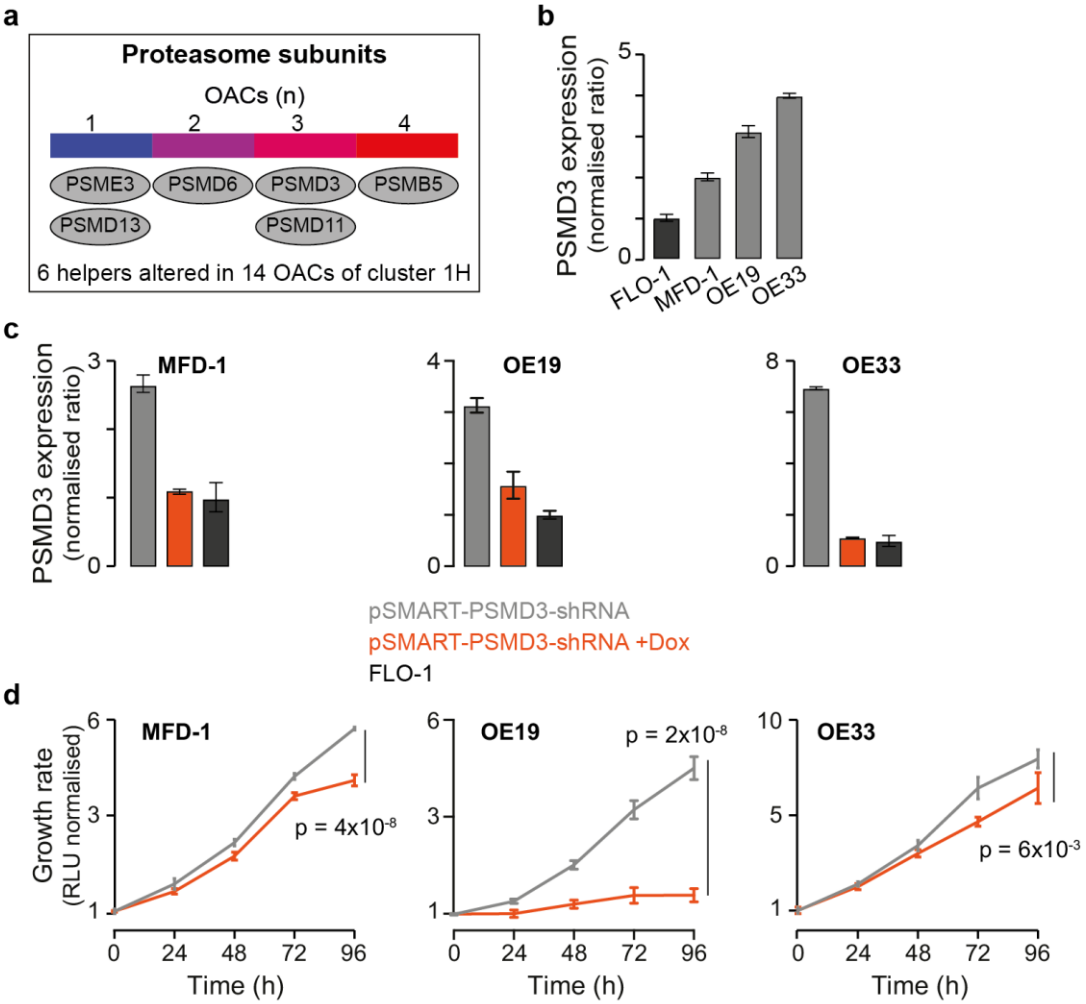


Figure 4.14. OAC cell dependence on *PSMD3* alteration. **(a)** Heatmap of proteasome subunits predicted as helpers in 261 OACs. **(b)** PSMD3 basal mRNA expression levels in FLO-1, MFD-1, OE19 and OE33 cells. Expression was relativized to β-2-microglobulin and normalised to FLO-1 cells. **(c)** PSMD3 expression levels in MFD-1, OE19 and OE33 after transduction with a lentiviral vector carrying an inducible shRNA against *PSMD3*. Expression was assessed in absence of doxycycline and after 96 hours of doxycycline treatment, relativized to β-2-microglobulin and normalised to FLO-1 cells. **(d)** Proliferation curves of MFD-1, OE19 and OE33 cells with or without doxycycline treatment to reduce PSMD3 expression to levels comparable to those of FLO-1 cells. For all proliferation assays, at least two biological replicates were performed, each with four technical replicates. Proliferation was assessed every 24 hours and each time point was normalised to time zero. Mean values at 96 hours were compared by two-tailed Student's t-test.

4.4 Discussion

In this part of my thesis, I utilised cancer helpers that were identified by sysSVM to stratify OACs and describe recurrently perturbed biological processes with tumorigenic potential. Instead of working at a gene level, pathway-level analysis substantially reduces the heterogeneity of the molecular landscape in OAC and allows the refinement of subgroups of patients with distinct molecular and clinical characteristics. I showed that when only known cancer genes were considered, patient groups were dictated by the mutational status of *TP53* and a few other known driver genes (Figure 4.2A and 4.2B). This is possibly a by-product of many years of research on these genes, which led to their representation in a high number of biological pathways. Instead, when cancer helpers were considered, the mutational status of *TP53* was irrelevant to the clustering of OACs. This observation suggests that the two clustering approaches capture different types of molecular characteristics. Closer examination of helper-derived OAC clustering revealed novel dysregulated pathways that, upon experimental validation, proved to play a role in tumorigenesis. Apart from novel pathways, helpers contributed to several well-known tumorigenic pathway, including intracellular signalling, cell cycle control checkpoints and DNA repair. Interestingly, while known driver genes tend to encode upstream players in these pathways, helpers were found to encode downstream effectors. For example, we found several Rho GTPase effectors (Figure 4.13A) and genes downstream of previously reported OAC drivers in the TLR cascades (Appendix Table 7.2).

Analysis of pathways disrupted by helpers in each OAC group (Figure 4.2A), revealed many functionally related processes. For example, clusters 1H and 3H shared perturbations in intracellular signalling and clusters 2H and 4H

shared perturbations in pathways involved in cell cycle regulation, namely S-phase entry and DNA replication. This suggests that tumorigenic processes can be somatically perturbed by more than one mechanism. Consistent with the perturbation of cell cycle regulation, clusters 2H and 4H brought together the most genomically unstable tumours. Moreover, several clusters showed enrichment in specific known driver alterations.

To establish the tumorigenic role of helpers in OAC, I collaborated with two other members of the Ciccarelli lab to experimentally validate representative helpers. By experimentally mimicking the amplification of *E2F1* (representative of cluster 2H) and *MCM7* (representative of cluster 4H), we showed increased proliferation in OAC cells when these genes were overexpressed (Figure 4.9C). We also provided evidence that *E2F1* increased proliferation by promoting S phase entry (Figure 4.10A), while overexpression of *MCM7* resulted in a reduction of MCM complex loading onto chromatin (Figure 4.10B and 4.10C), maybe due to a stoichiometric imbalance of complex subunits. This may indicate that MCM7 promotes cell growth through a separate mechanism besides its function in the MCM complex. For example, MCM7 interacts with the tumour suppressor protein Rb, a well-characterised inhibitor of E2F1 (Sterner et al. 1998). It is therefore possible that MCM7 overexpression may sequester Rb away from E2F1, thereby promoting E2F1-mediated cell cycle progression. Alternatively, it is also possible that stoichiometric balance of MCM complex was not reached in our experiments due to the duration of the experiment. Of note, preliminary evidence from the pan-cancer analysis suggests that stoichiometric balance of the MCM complex is reached in cancer cells upon overexpression of one member of the complex (Figure 4.8B). Finally, reducing the levels of *MCM7* expression in cells with high basal expression

levels decreased cell proliferation, suggesting not only a tumorigenic role of *MCM7* in OAC, but also a dependence of cancer cells on its high expression.

Apart from *E2F1* and *MCM7*, we also confirmed the cancer promoting role of very rare helpers, such as *ABI2*, *NCOR2* and *PAK1* that were altered between 1% and 4% of OACs. Therefore, irrespective of the frequency, helpers have a substantial impact on the progression of the cancer where their alteration occurs. This may indicate new, patient-specific gene dependencies and suggest possible stratifications that could inform the selection of targeted treatments. For example, 14 samples of cluster 1H have alterations of several proteasome subunits. Experimentally reverting the expression of the proteasome subunit PSMD3 to diploid levels resulted in reduced cell growth in three different OAC cell lines (Figure 4.14D). This indicates that OACs depend on helper alterations and are vulnerable to their inhibitions. Interestingly, proteasome inhibition has been shown to have a synergic effect in combination with ERBB2 inhibitors (Issaenko et al. 2012). Since *ERBB2* is also significantly altered in cluster 1H (Figure 4.2B), a combined therapy may be beneficial to patients in this cluster.

In summary, we provide one of the first attempts to extend the discovery of acquired perturbations contributing to cancer beyond those of recurrent drivers. Additional efforts are required to fully exploit the potential of these approaches to offer a more comprehensive view of the molecular mechanisms behind cancer and to guide novel clinical interventions.

Chapter 5. Discussion and Future Directions

5.1 Introduction

This thesis describes the development of the sysSVM algorithm for the identification of patient-specific cancer drivers. sysSVM combines supervised machine learning and systems biology in order to prioritise genes with tumorigenic potential amongst all genes with putative damaging alterations within a tumour. This is of particular interest for precision medicine approaches whose scope is to tailor therapeutic interventions to the mutational landscape of each tumour (Mathur and Sutton 2017; Jiang and Wang 2010; Xue and Wilcox 2016). In fact, curative treatments in several types of cancer have been hindered by widespread inter-tumour heterogeneity (Vogelstein et al. 2013), which makes targeted therapies challenging as few recurrent alterations are shared across samples. One such cancer type is OAC, in which very few alterations exceed 10% of recurrence across patients (Secrier et al. 2016). Throughout this thesis, I argued that the reason why very few cancer driver genes have been described in OAC, is because current methods, albeit statistically complex, use recurrence as a direct or indirect measure to assess significance and positive selection (Lawrence et al. 2013; Dees et al. 2012). SysSVM is instead focused on the identification of cancer drivers in the individual patients by investigating their systems-level and molecular properties.

The application of sysSVM to OAC provided insights into the rare and patient-specific drivers operating in this cancer type and the processes that they perturb. It also allowed the identification of sub-groups of patients with molecular alterations in shared pathways and processes. In this chapter, I summarise the main features of sysSVM, highlighting its limitations and extensions and commenting on the future application and clinical relevance.

5.2 SysSVM features are derived from systems-level and molecular properties of known cancer driver genes

SysSVM builds on previous attempts to define properties of cancer driver genes (D'Antonio and Ciccarelli 2013; An et al. 2016; Rambaldi et al. 2008; D'Antonio and Ciccarelli 2011) and subsequently uses these features to identify drivers in individual tumours (D'Antonio and Ciccarelli 2013). Initial studies, aiming to characterise driver genes, primarily concentrated on the human protein-protein interaction network and highlighted that proteins encoded by cancer genes are highly connected and central (Rambaldi et al. 2008; Jonsson and Bates 2006). This has been interpreted as a sign of fragility of cancer genes, as damaging alterations in these genes are likely to affect multiple biological processes within a cell (Ciccarelli 2010).

In addition to network-related properties, cancer genes share other systems-level properties, as well, that were included in the development of sysSVM and are summarised in Table 3.2. For example, cancer genes are expressed in a higher number of tissues than the rest of human genes and their function is regulated by a high number of miRNAs, indicating a tight control of their expression (An et al. 2016). Cancer genes have emerged in two rounds during evolution: tumour suppressors are mostly ancient genes with an origin in prokaryotes, while oncogenes were mostly acquired in metazoans (Domazet-Lošo and Tautz 2010). This denotes that damaging alterations in tumour suppressors impair mostly basic cell functions, while those in oncogenes affect mostly regulatory functions. Moreover, based on data presented in chapter 1, I found that cancer genes tend to be located in open-chromatin regions of human genome (Figure 2.3). In addition to systems-level properties, several features of sysSVM were derived from sequencing data (summarised in Table 3.2) to

restrict predictions to genes with putative damaging alterations. These features included damaging point mutations, indels, copy number alterations and structural variants.

Despite the wealth of information that the above-mentioned features provide sysSVM with, there are some limitations. The positive pairwise correlation that was observed between features, such as the number of exonic mutations and the number of damaging mutations (Figure 3.5A), denotes a level of redundancy that needs to be further investigated. Although a "ratiometric" version of these features has been used before to identify driver genes (Tokheim et al. 2016), their contribution to the predictive models of sysSVM needs to be explored in multiple cancer types. In addition, features derived from the protein-protein interaction network, such as degree and betweenness, were also correlated. However, this is a known/established correlation, as several hubs of the protein-protein interaction network are also central proteins. The fact that not all central proteins are hubs justified the usage of both these features in sysSVM.

During the development of sysSVM, I attempted to incorporate as many features as possible. Therefore, I performed an empirical selection of features exhibiting statistical significance between known cancer genes and the rest of human genes. However, there are other methods for feature selection and assessment (Anthony and Ruther 2007), whose application was not investigated. For instance, dimensionality reduction could be utilised before training to identify features enriched in outliers in the training set (Mwangi, Tian, and Soares 2014) and subsequently exclude them.

Moreover, the molecular properties used in this work were limited by the availability of data. Many other types of data, such as expression, methylation,

post-translational modification and miRNA regulation data could be utilised by sysSVM. With the growing availability of additional molecular data, a desirable future direction of sysSVM will be to examine whether additional molecular properties can reveal novel cancer drivers. For instance, additional data might include tumour-specific transcriptional regulation via epigenomic mechanisms (Villanueva et al. 2015; Chatterjee, Rodger, and Eccles 2018) or post-transcriptional modifications (Jewer, Findlay, and Postovit 2012). The former has been observed in the case of the proto-oncogene *KIT*, a type III receptor tyrosine kinase, which has been found overexpressed without gene amplification or mutation in paediatric renal tumours (Jones et al. 2007). Post-transcriptional modifications might involve alternative splicing to modulate protein structure through inclusion or skipping of exon(s). This has been previously reported for CCN proteins, whose alternatively spliced transcripts have been identified in multiple malignancies (Jewer, Findlay, and Postovit 2012). In addition, incorporation of expression data could facilitate modelling of complex scenarios contributing to tumorigenesis, such as tumour-specific RNA editing (Peng et al. 2018). To this end, direct variant calling from RNA-sequencing data and subsequent comparison with exome variants and normal tissue could be implemented to identify variants present only in the transcriptome of the tumour. RNA editing has been considered an overlooked source of cancer mutations and it has been shown to generate coding mutations and neoantigens in multiple cancer types (Nishikura 2016; Peng et al. 2018; Ben-Aroya and Levanon 2018; Roth et al. 2018). Although currently there are no known drivers with RNA editing mutations, recent studies suggest that RNA editing mutations can be driver mutations (Han et al. 2015), and therefore this could be a feature worth including in future versions of sysSVM

5.3 SysSVM is based on one-class support vector machines and predicts patient-specific cancer drivers

In several areas of biology, massive labelled data cannot be easily collected and, therefore, binary learning (i.e machine learning with both positive and negative observations) is challenging, as there are no negative observations that sufficiently represent the “negative” feature space. Although it is possible to use unlabelled observations as an opposing class to the positive set (Mordelet and Vert 2010; Xiaoli Li and Liu 2003; Chawla and Karakoulas 2005), this might lead to unbalanced training sets. This problem is even more prominent in the case of predicting cancer genes, as cancer drivers can be context dependent, and the same gene may or may not be a driver depending on the tissue (Schneider et al. 2017).

Because no available true negative set could be used for training, sysSVM employed a one-class support vector machine framework (Scholkopf et al. 2001), in which the density of known cancer genes in the feature space is modelled. Despite the robust theoretical background of one-class support vector machines, several other algorithms, such as isolation forest (Liu, Ting, and Zhou 2008) and local outlier factor (Breunig et al. 2000), have been developed for one-class classification. Such algorithms are directed towards outlier detection in an effort to define the high-density regions of the feature space. This is particularly relevant to one-class classification, as the absence of true negatives might lead to a wide decision boundary that is prone to false positives. Isolation forests operate on a fundamentally different approach than the one of one-class support vector machines. Rather than modelling positive observations, isolation forests model anomalies in the training data based on the fact that anomalies are the minority in the training data and they have

feature values that deviate significantly from the positive observations (Liu, Ting, and Zhou 2008). Using a similar approach, in isolation forests local outlier factor algorithm uses density-based clustering to indicate which training observations have high degree of outlier-ness (Breunig et al. 2000). Although one-class support vector machines is the most robust and widely used algorithm, the implementation of isolation forests or local outlier factor for the identification of patient-specific cancer drivers merits further investigation. Future development of sysSVM could be directed towards the combination of alternative algorithms to assess whether it would potentially lead to a more robust decision boundary around the high-density regions of the feature space than each algorithm on its own.

Another limitation of sysSVM is that it does not explicitly model cancer genes as tumour suppressors or oncogenes, but it only does it in an indirect way. Since several systems-level properties exhibit differences between tumour suppressors and oncogenes, this information is already embedded in the data used in sysSVM. But future implementations of the algorithm could leverage on this distinction more explicitly and utilise a pre-defined feature weighting vector (Zhang et al. 2009) of systems-level properties. First, the tumour suppressor-specific features, such as ancient evolutionary origin or the absence of duplicated copies in the human genome (Repana et al. 2018), could be assigned with a higher weight during training, than the rest of the features. This would force sysSVM to assign higher score to genes that resemble tumour suppressors. Repeating the same process for the oncogene-specific features would result in two scoring systems, whose top-scoring genes would correspond to predicted tumour suppressors and oncogenes. In addition to feature weighting, a gene-specific refinement of the molecular properties, either

before or after prediction, could also be implemented to ensure that oncogene and tumour suppressor predictions harbour activating and damaging alterations, respectively.

5.4 Patient-specific cancer drivers reveal widespread perturbation of biological processes in OAC

In chapter 3, I applied sysSVM to a cohort of 261 OACs to identify rare and patient-specific cancer drivers. OAC was considered a very good test case for the application of sysSVM, as the high inter-patient heterogeneity in this cancer type suggests that rare drivers might be involved in tumour development. In particular, I hypothesised that alongside the critical role of recurrent and well-known drivers, complementary alterations of several other genes might contribute to cancer development. These cancer “helper” genes were identified using sysSVM and their relevance to OAC was examined using pathway analysis and experimental validation. However, as mentioned previously, sysSVM is not designed to identify only cancer helpers, but to predict cancer drivers in general. In the cohort of 261 OACs, known cancer genes were used for training of sysSVM and, therefore, they were not part of the prediction set. For this reason, the new predictions did not contain any known drivers. Nonetheless, several of the predicted helpers may have a driver role in cancer cells that is as strong as the one of the known drivers. The potential of sysSVM to identify known drivers, as well as helpers, was demonstrated when it was applied to two validation cohorts (Figure 3.10), confirming that trained models can successfully identify known drivers.

SysSVM highlighted 952 top-scoring genes as helpers in OAC (Appendix Table 7.2). The majority of these genes (~ 80%) were implicated in amplification events that led to overexpression (Figure 3.11B), suggesting that their amplification has a possible functional role in OAC. The number of predicted helpers is significantly higher than that of other methods even in pan-cancer cohorts (Bailey et al. 2018; Lawrence et al. 2013). This is due to the fact that approximately 60% of these genes were rare or patient-specific, with only few genes being altered in more than 5% of OACs. However, the recurrence of helpers might increase if more than ten high-scoring genes will be considered. Intersection of helpers with two lists of false positive cancer drivers (including more than 500 genes) confirmed the low false positive rate of sysSVM, as only 44 helpers (4.6% of the total) were previously identified as possible false positives.

Helpers predicted in multiple OACs did not always harbour the same type of alteration. This adds to the growing body of evidence suggesting that different alterations in drivers might not have the same impact (Pardo and Godzik 2015). One of the most well-known examples is that of the two most common mutations in *PIK3CA*, E545K and H1047L, which contribute to tumorigenesis through different mechanisms (Hao et al. 2013). Similar observations have been reported for *KRAS* (Garassino et al. 2011), *EGFR* (Porta-Pardo et al. 2015) and *TP53* (Porta-Pardo et al. 2015). These observations indicate that sysSVM could be implemented using only systems-level properties as predictive features and utilise the molecular properties (i.e. damaging alterations) only after prediction to refine the predicted drivers. Such an implementation merits further investigation and it could be facilitated by

extension of sysSVM to additional cancer types to assess its feasibility and potential biases.

One line of investigation that has not yet been explored is the relationship between the molecular landscape and the ranking of helpers in individual OACs. For example, I observed that several genes that harboured damaging alterations in multiple OACs were amongst the top-scoring predictions in only a fraction of them. This is due to the fact other high-scoring helpers are present in these samples. Although by design sysSVM score incorporates information on the number of altered genes within each sample (see equation 2.10), the extension of sysSVM to additional cancer types will enable further investigation of these genes. One possibility will be to explore whether genes with specific sysSVM features that have low weight in all kernels tend to be implicated in these discrepancies more frequently than others. The investigation of these discrepancies is also relevant to the application of sysSVM in new cohorts with already-trained models. The analysis that was presented in chapter 4 to assess whether sysSVM successfully recognises known cancer drivers may provide insights on the aetiology of ranking discrepancies, as not all known cancer drivers were amongst the top-scoring genes.

To examine the mechanisms through which helpers might contribute to tumorigenesis, I performed a gene set enrichment analysis using almost 2,000 pathways. This revealed that several helpers converged towards the perturbation of well-known tumorigenic pathways to which known driver genes also contribute. These pathways were related to intracellular signalling, cell cycle control, apoptosis and DNA repair, and were associated with the most recurrently altered known drivers (*TP53, CDKN2A, MYC, ERBB2, SMAD4, CDK6, KRAS*). This is a confirmation of the reliability of sysSVM predictions, as

close interactors of known drivers might be true positive predictions of cancer drivers in OAC. However, the uncertainty of this observation needs further investigation. This might represent an important refinement of my findings, as it would provide a better indication of the non-random membership of helpers to these pathways and exclude possible false positive findings. Such an extension could be implemented via multiple randomisation steps through which random sets of helpers would be analysed progressively for their membership to these pathways.

Apart from the known tumorigenic pathways, in chapter 4 I showed that helpers perturb several pathways, whose contribution to OAC is not fully understood. For example, perturbations in the TLR signalling cascades were found dysregulated in more than 10% of OACs. Shortly after our observation, an independent study that used a pathway-oriented algorithm to find significantly altered biological processes in OAC confirmed these findings (Fels Elliott et al. 2017). TLRs are pattern recognition receptors, which are involved in innate immune system and in the interaction of cells with microbiota (Akira and Takeda 2004). Initial findings reported somatic mutations in *TLR4* gene in multiple solid tumours, including OAC (Hold et al. 2007; Kurt et al. 2016; Fels Elliott et al. 2017; Dulak et al. 2013). My results suggest a wider perturbation of TLR signalling cascades in OAC, which might lead to a broader dysregulation of inflammatory signalling. Future research will further elucidate the mechanisms via which tumour cells utilise TLRs to promote their survival and evade anti-tumour immune responses in OAC. Furthermore, it is worth noting that the specific genes that were predicted by sysSVM, were not identified by alternative approaches, such as that of Fells-Elliot and colleagues (Fels Elliott et al. 2017).

This finding highlights that sysSVM complements current state-of-the-art methods for cancer driver discovery.

In addition to the TLR cascades, sysSVM predicted Rho GTPase activity as another major contributor to OAC pathogenesis in approximately 20% of tumours. Recurrent mutations in members of the RAC1 GTPase pathway, such as *ELMO1* and *DOCK2*, were previously identified in OAC, but they were found related to increased invasive capacity, rather than cell proliferation (Dulak et al. 2013; Weaver, Ross-Innes, and Fitzgerald 2014). Our results suggest that alterations in members of Rho GTPase activity and, in particular, of Rho GTPase effector pathways in OAC affect cell proliferation (Figure 4.13). This indicates that the implementation of methods oriented towards patient-specific detection of cancer drivers can indeed reveal novel perturbations in the level of processes that are not detectable when individual recurrently altered driver genes are examined.

5.5 Cancer helpers highlight six OAC patient sub-groups with putative therapeutic implications

Targeted therapies in OAC lag behind those in other cancer types and even recent clinical trials reported disappointing or inconclusive results (Kopp and Hofheinz 2016; Woo, Cohen, and Grim 2015; Young and Chau 2016). This is at least partially due to the incomplete description of the driver alterations in OAC that leads to an inaccurate molecular stratification of patients (Weaver, Ross-Innes, and Fitzgerald 2014). The comprehensive characterisation of helpers using sysSVM allows a fine-mapping of drivers in each OAC and favours the refinement of subgroups of patients using their molecular

landscape. In chapter 4, I described six OAC clusters based on helpers and confirmed that patient stratification using only known drivers is confounded by the mutational status of *TP53* and a few other known drivers (Figure 4.2). Conversely, patient stratification using helpers highlighted sub-groups of OACs that were brought together by damaging alterations of different helpers in the same pathways. Interestingly, several of these clusters were associated with specific background alterations of major drivers such as *RECQL4*, *RARA*, *MYC*, *SMARCE1* and *ERBB2* in cluster 1H, *GNAS*, *SS18L1*, and *FHIT* in cluster 2H and *TRAPP* and *CDK6* in cluster 4H.

Further experimental validation and analysis is needed to confirm the clinical relevance of these patient clusters. The findings in this thesis provide several lines of evidence to guide future studies, as several associations with potential clinical relevance were described in each cluster. For example, a profile of increased genomic instability was observed in patients of clusters 2H and 4H (Figure 4.5). Given the connection of genomic instability, neoantigen generation and tumour infiltration from immune cells (Yaghmour et al. 2016; Chan et al. 2018), this analysis could highlight a sub-group of OACs that might benefit from immunotherapy. As more expression data become available for this cohort, tumour infiltrates and PD1/PD-L1 expression could be quantified and linked back to the molecular stratification using helpers. Moreover, clusters 1H and 4H exhibited high exposure to *BRCA1* and *BRCA2*-related mutational signatures, which may render them sensitive to PARP inhibitors, which have shown promising results (Lord and Ashworth 2017).

The high number of helpers predicted using sysSVM made the experimental validation of all of them unfeasible. However, we selected representative genes for experimental validation based on their importance in

defining patient clusters or their rarity. Validation of rare helpers was an important step of this project as our hypothesis was that rarely altered genes contribute to recurrent perturbations of biological processes. A further validation of helpers could be performed on cancer cells with matched background alteration of known drivers (e.g. *MYC* or *ERBB2* for helpers in cluster 1H) to confirm their helper role in a specific molecular background, such as that described in figure 4.2.

One main question that arises is how sysSVM could be implemented in clinical practice. Apart from the refinement of patient sub-groups that may provide insights into therapy, practical application of sysSVM for classification of individual patients and prediction of drivers and helpers could utilise the already trained models that were presented in this thesis. Although the number of OACs used in this work is about a third of those that have been sequenced so far by multiple studies, further analysis is needed to define the appropriate size of the training set in a clinical setting and quantify the contribution of new training observations. In addition to that, the training set of sysSVM could be subject to regular updates of its training set as new patients come into clinic. To this end alternative algorithms to the one used here (i.e. grid search) could be used to efficiently optimise the parameters of each kernel (Manurung, Mawengkang, and Zamzami 2017; Tsuruoka, Tsujii, and Ananiadou 2009). A paradigm of a rolling update of the training set might lead to better description of parts of the feature space that are not well represented in the current implementation. As a result, sysSVM will be able to analyse the molecular landscape of patients with non-recurrent or not previously seen drivers or helpers. Once trained, sysSVM models can be applied to individual patients even using personal computers in a clinical setting, as the prediction step is not computationally expensive.

5.6 Conclusion

The work presented in this thesis provides new insights to the field of cancer driver discovery. The development and application of sysSVM confirmed the ability of one-class support vector machines and systems-level properties to predict cancer drivers. Our findings provided new insights to the understanding of OAC biology, as previously-unknown genes contributing to OAC development were revealed by applying sysSVM. Several of these genes, such as *E2F1* and *MCM7*, were experimentally validated in OAC cell lines and were found to significantly increase the proliferation rate of cancer cells. Another important finding is that sysSVM top-scoring predictions converged towards the perturbation of the same biological processes, which were further divided into well-known and novel tumorigenic processes. Participation of helpers in well-known tumorigenic pathways denotes that they are true positive predictions as they are primary interactors of known cancer drivers. A further important finding has been the identification of novel and putative targetable processes operating in OAC, such as TLR signalling cascades and proteasome activity. This finding may also inform patient stratification, as targeted therapies are in their infancy in OAC and initial results have been disappointing. Patient clustering using helpers revealed six patient sub-groups with distinctive molecular profiles and potential implications for clinical practice. Finally, sysSVM exemplifies cancer driver discovery without the use of recurrence of gene alterations. Such methods are of interest especially in cancer types with widespread inter-patient heterogeneity, such as OAC. The application of sysSVM in other cancer types could be used to guide testable hypothesis regarding the activity of cancer drivers in patients with unknown genetic determinants.

Chapter 6. Materials and Methods

6.1 Cohort description

Samples included in this study provided by the Oesophageal Cancer Clinical And Molecular Stratification (OCCAMS) consortium, which was incorporated into the International Cancer Genome Consortium (ICGC). All patients gave individual informed consent. OAC samples were obtained during surgical resection or by biopsy at endoscopic ultrasound, as previously reported (Secrier et al. 2016). Briefly, normal squamous oesophageal samples, resected at least 5 cm away from the tumour region, or blood were used as germline reference. All tissue samples were snap-frozen after collection. Prior to DNA extraction, the cellularity of hematoxylin-and-eosin-stained sections of the samples were assessed by two expert pathologists. Cancer samples with cellularity ≥ 70% were submitted for whole genome sequencing (WGS). DNA was extracted from frozen oesophageal tissue and from blood samples as previously described (Secrier et al. 2016). A total of 261 cases (matched tumour–normal) were sequenced and used in this study (Table 3.3). A smaller cohort of 18 OACs, which were part of the 261 cases, was used during the development phase of sysSVM.

6.2 Annotation of molecular properties

Data on somatic single nucleotide variations (SNVs), small insertions and deletions (indels), copy number variations (CNVs), structural variations (SVs), and mutational signatures for 261 OACs were obtained from ICGC and analysed as previously described (Secrier et al. 2016). Briefly, SNVs and indels

were called using Strelka v.1.0.13 (Saunders et al. 2012) and subsequently filtered to obtain high quality calls. Filters are summarised in Table 6.1. For CNVs, the absolute copy number for each genomic region was obtained from ASCAT-NGS v.2.1 (Van Loo et al. 2010) after correction for tumour content, using read counts at germline heterozygous positions as derived from GATK v.3.2-2 (McKenna et al. 2010). To account for the high number of amplifications occurring in OAC, copy number gains were corrected by the ploidy of each sample as estimated by ASCAT-NGS. A gene was assigned with the copy number of a CNV region if at least 25% of its length was contained in that region. SVs (gene translocations, inversions, insertions) were identified from discordant read pairs using Manta (Chen et al. 2016) after excluding SVs that were also present in more than two normal samples of a panel of 15 oesophagus and 50 blood samples. For the validation cohort from TCGA, SNVs, indels, and CNVs were derived from level 3 TCGA annotation data of 86 OACs (https://portal.gdc.cancer.gov/projects/TCGA-ESCA). In the case of 21 OACs from a previous study (Nones et al. 2014), SNVs, indels, and CNVs were called as described for the ICGC samples. The distribution of variant allele frequency of SNVs and indels across all samples was used to remove outliers likely indicating sequencing or calling artefacts. Variants with <10% frequency and indels longer than five base pairs were also removed. For CNVs, genomic regions were considered as amplified or deleted if their segment mean was higher than 0.3 or lower than -0.3, respectively. A gene was considered as amplified or deleted if at least 25% of its length was contained in a CNV region and the resulting copy number (CN) was estimated as:

$$CN = 2 \times 2^{segment\ mean}$$

No SV data were available for the validation cohorts.

Since only genes with predicted damaging alterations were used as input for sysSVM, further annotation for the variant damaging effect was performed. stopgain, stoploss, frameshift, nonframeshift, nonsynonymous, and splicing SNVs and indels were annotated using ANNOVAR (December 2015) (Wang, Li, and Hakonarson 2010). All truncating alterations (stopgain, stoploss, and frameshift mutations) were considered as damaging. Nonframeshift and nonsynonymous mutations were considered as non-truncating damaging alterations, if predicted by:

i) at least five of seven function-based methods: (SIFT (Kumar, Henikoff, and Ng 2009), PolyPhen-2 HDIV (Adzhubei et al. 2010), PolyPhen-2 HVAR (Adzhubei et al. 2010), MutationTaster (Schwarz et al. 2010), MutationAssessor (Reva, Antipin, and Sander 2011), LRT (Chun and Fay 2009) and FATHMM (Shihab et al. 2013)), or

ii) two out of three conservation-based methods (PhyloP (Pollard et al. 2010), GERP++ RS (Davydov et al. 2010), SiPhy (Garber et al. 2009)),

using the scores from dbNSFP v.3.0 (Liu et al. 2016).

Splicing modifications were considered as damaging if predicted by at least one of the two ensemble algorithms implemented in dbNSFP v3.0. Putative gain of function alterations were predicted using OncodriveClust (Tamborero, Gonzalez-Perez, and Lopez-Bigas 2013) with default parameters and a false discovery rate of 10%. The transcript lengths to estimate mutation clustering were derived from the refGene table of UCSC Table Browser (https://genome.ucsc.edu/cgi-bin/hgTables). Gene gains, homozygous losses, translocations, inversions, and insertions were always considered as putative damaging alterations.

Table 6.1. Summary of filters applied to SNV calls from Strelka. Analysis pipeline was applied as previously reported (Secrier et al. 2016).

Filter	Description
DistanceToAlignmentEndMedian	The median shortest distance of the variant position within the read to either aligned end is less than 10
DistanceToAlignmentEndMAD	The median absolute deviation of the shortest distance of the variant position within the read to either aligned end is less than 3
LowMapQual	The proportion of reads at the variant position with low mapping quality (less than 1) is greater than 10%
MapQualDiffMedian	The difference in the median mapping quality of variant reads (in the tumour) and reference reads (in the normal) is greater than 5
VariantMapQualMedian	The median mapping quality of variant reads is less than 40
VariantBaseQualMedian	The median base quality at the variant position of variant reads is less than 30
VariantAlleleCount	The number of variant-supporting reads in the tumour is less than 4
VariantAlleleCountControl	The number of variant-supporting reads in the normal is greater than 1
StrandBias	The strand bias for variant reads covering the variant position, i.e. the fraction of reads in either direction, is less than 0.02, unless the strand bias for all reads is also less than 0.02

Repeat	The length of repetitive sequence adjacent to the variant position, where repeats can be 1-, 2-, 3-, or 4-mers, is 12 or more
SNVCluster50	The largest number of variant positions within any 50 base pair window surrounding, but excluding, the variant position is greater than 2; variant positions are those in which the number of alternate allele is supported by at least 2 reads and at least 5% of all reads covering that position
SNVCluster100	The largest number of variant positions within any 100 base pair window surrounding, but excluding, the variant position is greater than 4; variant positions are those in which the number of alternate allele is supported by at least 2 reads and at least 5% of all reads covering that position

6.3 Annotation of systems-level properties

Protein sequences from RefSeq v.63 (Pruitt et al. 2014) were aligned to the human reference genome assembly GRCh37 to define unique gene loci, as previously described (An et al. 2016). The length of the longest coding sequence was taken as the gene length. Genes aligning to more than one gene locus for at least 60% of the protein length were considered as duplicated genes (Rambaldi et al. 2008). Data on human ohnologs (gene duplicates retained after whole genome duplications) were collected from Makino et al., 2013 (Makino, McLysaght, and Kawata 2013). The number of protein domains was derived from CDD (Marchler-Bauer et al. 2013). The gene chromatin state

(a measure of the degree to which the DNA is densely packed) based on Hi-C experiments (Lieberman-Aiden et al. 2009) was retrieved from the covariate matrix of MutSigCV v1.2.01 (Lawrence et al. 2013). Chromatin state of each gene (measured in K562 cells) was ranging approximately from -50 (very closed) to +50 (very open). Data on protein-protein and miRNA-gene interactions, gene evolutionary origin and gene expression were retrieved as described in An et al., 2016 (An et al. 2016). Briefly, human protein-protein interaction network was rebuilt from the integration of BioGRID v.3.4.125 (Chatr-aryamontri et al. 2017); MIntAct v.190 (Orchard et al. 2014); DIP (April 2015) (Salwinski et al. 2004); HPRD v.9 (Keshava Prasad et al. 2009); the miRNA-gene interactions were derived from miRTarBase v.4.5 (Hsu et al. 2014) and miRecords (April 2013) (Xiao et al. 2009); gene evolutionary origin was assessed as described in D'Antonio et al., 2011 (D'Antonio and Ciccarelli 2011) using gene orthology from EggNOG v.4 (Powell et al. 2014); and gene expression in 30 normal tissues was retrieved from GTEx v.1.1.8 (Mele et al. 2015). Except gene length, duplication and ohnologs, all other systems-level properties had missing information for several human genes (Table 3.2). To account for this, median imputation for continuous properties and mode imputation for categorical properties were implemented. Specifically, for each property median or mode values were calculated for known cancer genes and the rest of mutated genes. All missing values were replaced by their corresponding median or mode values.

6.4 Application of sysSVM to 261 OACs

The three steps of sysSVM were applied to 261 OACs (Figure 3.3A). In step 1, all 34 features derived from molecular and systems-level properties

(Table 3.2) were mapped to the 17,078 altered genes in the cohort. Each feature was scaled to zero mean and unit variance to correct for the different numerical ranges across them. In step 2, 476 known cancer genes with damaging alterations (Appendix Table 7.1) were used as a set of true positives for model selection. To optimise the parameters of the four kernels (linear, radial, sigmoid and polynomial) a grid search using 10,000 iterations of a three-fold cross validation was performed. At each iteration, the 476 known cancer genes were randomly split into 2/3 (approximately 317 genes) that were used as a training set and 1/3 (approximately 159 genes) that was used as the test set. At each increment of 100 cross validation iterations, the four best models (one per kernel) were chosen based on the median and variance of the sensitivity distribution across all previous iterations of cross-validation. The selection of the 100 sets of best models from all 10,000 cross-validation iterations was repeated 5 times, where all iterations were randomly re-ordered. In step 3, the resulting 500 best models were trained with the whole training set and used to rank the remaining 16,602 unique genes in each patient. A score was calculated to combine the predictions form the four kernels and the genes that were not expressed in normal oesophagus, according to GTEx annotation, were excluded. These produced 500 lists of top 10 genes. Out of 500 best models, 38 had a unique set of parameters resulting in 24 unique lists of top 10 genes (Table 3.4). These 25 lists ranged between 898 and 952 genes, with a core set of 598 genes shared across all of them. The most frequent top 10 list occurred 207 times (952_A, 41.4%, Table 3.4). It was followed by 952_B (32.2%, 161 times) and 951_A (8.6%, 43 times). These three lists accounted for 82.2% of the 500 sets of top 10 genes, they shared 950 genes and were predicted by models differing in only one parameter (gamma in the polynomial

kernel, Table 3.4). Furthermore, the most frequent list was always predicted by the same set of best models. Therefore, 952_A represented a robust set of prediction and was considered as the final list of helper genes (Appendix Table 7.2).

6.5 Identification of perturbed processes and patient clustering

To identify the perturbed biological processes in OAC, both predicted cancer helper genes and known cancer driver genes were used. A manual revision of 476 known cancer genes that were altered in the ICGC cohort was performed and genes were considered as known drivers if:

i) their somatic alteration had been previously associated with OAC,

ii) they had a loss-of-function alteration and their tumour suppressor role had been reported in other cancer types (Vogelstein et al. 2013), or

iii) they had a gain-of-function alteration and their oncogenic role had been reported in other cancer types (Vogelstein et al. 2013).

The resulting 202 known cancer drivers and 952 cancer helpers were used for the gene set enrichment analysis using as reference Reactome v.58 (Fabregat et al. 2016), which is composed of 1,877 pathways and 10,131 genes. After excluding pathways in levels 1 and 2 of Reactome hierarchy and those with less than 10 or more than 500 genes, 1,155 pathways were retained. These contained 9,061 genes, including 155 known drivers and 648 helpers. Gene set enrichment was assessed using a one-sided hypergeometric test and the resulting *P* values were corrected for multiple testing using the Benjamini & Hochberg method (Appendix Tables 7.3 and 7.4). Enriched pathways within the sets of known drivers or helpers were subsequently used to cluster samples, taking into account the proportion of perturbed processes shared between

samples. The Jaccard index (*A*) was calculated by deriving the proportion of shared perturbed processes between all possible sample pairs as:

$$A_{ij} = \left|P_i \cap P_j\right| / \left|P_i \cup P_j\right|$$

where P_i *and* P_j are the perturbed processes in samples *i* and *j*, respectively.

Complete linkage hierarchical clustering using Euclidean distance between each row was performed on the resulting matrix. Clusters were visualised using ComplexHeatmap R package (Gu, Eils, and Schlesner 2016). To identify the optimal number of clusters, the median silhouette value of the samples between 3 and 20 clusters was measured as a measure of clustering robustness (Rousseeuw 1987).

6.6 Analysis of expression data

Purified total RNA was extracted from 92 out of the 261 OACs in the ICGC cohort and sequenced as described previously (Secrier et al. 2016). RNA sequencing reads were then aligned to human reference genome hg19 and expression values were calculated using Gencode v19. The summariseOverlaps function in the R GenomicAlignments package was used to count any fragments overlapping with exons (parameters mode=Union, singleEnd, invertStrand and inter.feature were set according to the library protocol, fragments=TRUE, ignore.strand=FALSE). Gene length was calculated as the number of base pairs in the exons after concatenating the exons per gene in non-overlapping regions. FPKM (Fragments Per Kilobase Million) were calculated for each gene as:

$$FPKM = \frac{gene\ read\ count}{(library\ size/1000000) \times (gene\ length/1000)}$$

Pan-cancer expression data for *KAT2A*, *KAT2B* and *E2F1* were downloaded from Xena browser (https://xenabrowser.net/).

6.7 Experimental validation

6.7.1 Materials

Antibodies and dyes

- The dye Alexa Fluor™ 488 NHS Ester (Succinimidyl Ester) used in barcoding was supplied from Thermo Fisher
- The full list of antibodies is provided below

Table 6.2. List of antibodies used in this study.

Test Antibody	**Class**	**Origin**	**Manufacturer**
Anti-human MCM7	monoclonal	mouse	Santa Cruz Biotechnology
Anti-human MCM3	polyclonal	rabbit	Bethyl Laboratories
Anti-mouse-Alexa Fluor 555	polyclonal	donkey	ThermoFisher Scientific
Anti-rabbit Alexa Fluor 555	polyclonal	donkey	ThermoFisher Scientific

Cell lines

Table 6.3. List of cell lines used in this study.

Cell line	**Origin**	**Source**
FLO-1	Human oesophageal adenocarcinoma	European Collection of Authenticated Cell Cultures

HEK293T	Human embryonic kidney	Francis Crick Institute Cell Services
MFD1	Human oesophageal adenocarcinoma	OCCAMS consortium
OE19	Human oesophageal adenocarcinoma	Francis Crick Institute Cell Services
OE33	Human oesophageal adenocarcinoma	European Collection of Authenticated Cell Cultures

Media and solutions

Table 6.4. List of media and solutions used in this study.

Product	**Supplier**
Dulbecco's Modified Eagle's Medium (DMEM)	Sigma-Aldrich
Dulbecco's phosphate buffered saline (DPBS)	Sigma-Aldrich
Ethylenediaminetetraacetic acid (EDTA)	Sigma-Aldrich
Fetal bovine serum (FBS)	Biosera
Geneticin (G-418) solution	Sigma-Aldrich
Glutamine Solution 200 mM	Sigma-Aldrich
HEPES buffer solution 1M	Sigma-Aldrich
Penicillin/Streptomycin solution 10,000 units penicillin and 10 mg streptomycin/mL	Sigma-Aldrich
Puromycin solution	Sigma-Aldrich

Kits and reagents

Table 6.5. List of kits and reagents used in this study.

Product	Supplier
Blasticidin	Sigma-Aldrich
Bovine Serum Albumin (BSA)	Sigma-Aldrich
Brilliant III Ultra-Fast SYBR Green QRT-PCR Master Mix	Agilent Technologies
CellTiter 96 Non-Radioactive Cell Proliferation Assay	Promega
CellTiter-Glo Luminescent Cell Viability Assay	Promega
Click-iT™ Plus EdU Alexa Fluor™ 647 Flow Cytometry Assay Kit	Invitrogen
cOmplete™ EDTA-free protease inhibitor cocktail tablets	Roche
DAPI	Invitrogen
Doxycycline	Sigma-Aldrich
EdU	Invitrogen
GeneArt Platinum Cas9 nuclease	Life Technologies
GeneElute mammalian total RNA miniprep kit	Sigma-Aldrich
Hexadimethrine bromide	Sigma-Aldrich
Lipofectamine CRISPR max	Life Technologies
Methanol (99.8%)	Sigma-Aldrich
Nuclease-free water	Promega, Wisconsin, USA
Q5 Hot Start High-Fidelity 2X Master Mix	New England Biolabs
RNAse A	Invitrogen
tracrRNA 69-mer	Sigma-Aldrich

Trypan Blue solution (0.4%) Sigma-Aldrich

Solutions prepared in-house

1. 1% BSA/PBS.

   Was prepared by dissolving 1g of BSA in 100 mL of PBS

2. Cytoskeleton buffer (CSK), provided by the Crick Lab stocks

   - 10mM HEPES-KOH pH7.9
   - 100mM NaCl
   - 3mM $MgCL_2$
   - 1mM EGTA (Ethylene glycol tetraacetic acid)
   - 300mM sucrose
   - 1% BSA
   - 0.2 Triton-X 100
   - 1mM DTT
   - cOmplete EDTA-free protease inhibitor cocktail tablets, Roche

Vectors

Table 6.6. List of vectors used in this study.

Vector name	**Insert**	**Selectable marker**	**Supplier**
pCMVHA E2F1	E2F1	neomycin	Addgene
pLX_TRC317	MCM7	puromycin	Sigma-Aldrich
pcDNA3.1+/C-(K)-DYK	PAK1	neomycin	Genscript
EZ-Tet-pLKO-Puro	emtpy	puromycin	Addgene
pcDNA3.1+/C-(K)-DYK	empty	neomycin	Genscript
GeCKO-tEf1aCas9Blast	Cas9*sp*	blasticidin	Sigma-Aldrich

6.7.2 Methods

All validation experiments were performed by Dr. Lorena Benedetti and Dr. Elizabeth Foxhall.

Cell culture

Overexpression and editing experiments were carried out using the FLO-1 oesophageal adenocarcinoma cell line. Gene knockdown experiments were performed on OE19, OE33 and MFD1 cells. FLO-1 cells were grown at 37°C and 5% CO_2 in DMEM supplemented with 2mM Glutamine, 10% FBS, 100 I.U./mL penicillin and 100 µg/mL streptomycin. OE19 and OE33 cells were grown in RPMI supplemented with 2mM Glutamine, 10% FBS, 100 I.U./mL penicillin and 100 µg/mL streptomycin. MFD1 cells were grown in DMEM supplemented with 2mM Glutamine, 10% FBS, 100 I.U./mL penicillin and 100 µg/mL streptomycin. All cells were maintained at 37°C and five per cent CO_2, validated by short tandem repeat analysis and routinely checked for mycoplasma contamination. For all experiments, cells were harvested in log phase growth using Trypsin-EDTA 1x solution, washed and resuspended in culture medium. To determine cell numbers and viability, cell suspension was mixed with known volume of 0.4% Trypan blue solution and then cells were counted using a haemocytometer.

Gene overexpression

The vectors pCMVHA E2F1 (Lukas et al. 1996), pLX_TRC317 and pcDNA3.1+/C-(K)-DYK were used to induce *E2F1*, *MCM7*, and *PAK1* overexpression, respectively (Table 6.6). Cells were transfected according to the manufacturer's protocol and selected with 0.5 ug/mL of either

G481/Geneticin (*E2F1*, *PAK1*) or Puromycin (*MCM7*). Empty vectors pcDNA3.1+/C-(K)-DYK and EZ-Tet-pLKO-Puro, carrying Neomycin or Puromycin resistance, respectively, were used as controls (Table 6.6).

RNA was isolated from transfected cells using GeneElute mammalian total RNA miniprep kit, as per manufacturer's instructions, and was used to assess gene overexpression via quantitative RT-PCR. Briefly, qRT-PCR was performed in a 20 ul reaction using predesigned SYBR green primers (Table 4.2 Primers, Sigma Aldrich) and Brilliant III Ultra-Fast SYBR Green QRT-PCR Master Mix (Agilent Technologies) for detection (10ul of 2× SYBR Green QPCR master mix, 0.8µl of primers and 7.2ul of dH20). Primers were diluted 1:10 from a 100uM stock. qRT-PCR was performed on the viiA7 Real-Time PCR System (Applied Biosystems) and settings were fixed: Hold Stage (95oC, 2 minutes), PCR cycles (40 in total) of Melt (95oC, 5 seconds), Anneal/Extension (63oC, 30 seconds).

The average expression level across triplicates (*e*) was relativized to the average expression level of β-2-microglobulin (*c*):

$$r = e - c$$

where *r* is the relative gene expression. The fold change (*fc*) between the relative gene expression after overexpression and the relative gene expression in the control condition (r_c) was calculated as:

$$fc = 2^{(r_c - rKD)}$$

Each sample was assessed in triplicate and each experiment was repeated in biological duplicate.

Gene editing

To induce gene knock-out (KO), the vector-free CRISPR-mediated editing approach was used (Lorena Benedetti et al. 2017). Initially, cells were transduced with lentiviral vector (Table 6.6) containing Cas9*sp* and blasticidin resistance marker. After 10 days of treatment with blasticidin (25 ug/ml), resistant cells were selected and Cas9*sp* expression was verified *via* PCR using specific primers (Table 4.2).

Consequently, Cas9*sp*-expressing cells were co-transfected using lipofectamine CRISPR max with a 69-mer tracrRNA, GeneArt Platinum Cas9 nuclease and three gene-specific crRNAs (Table 4.2). To avoid off-target editing, all crRNAs used were verified to map only the gene of interest with a perfect match and additional hits in the genome with at least three mismatches. Control cells were transfected with the same protocol but using three non-targeting control (NTC; Table 4.2) crRNAs.

Gene editing was confirmed with Illumina Miseq sequencing. Genomic DNA from edited cells was extracted and regions surrounding the targeted sites were amplified with primers containing Illumina adapters (Table 4.2) using Q5 Hot Start High-Fidelity 2X Master Mix in a 25ul reaction [120ng of DNA (DNA+H2O=10ul, 2X Q5 MM=12.5ul, 1.25 ul of each primer)]. DNA barcodes were added with a PCR reaction before pooling the samples for sequencing on Illumina MiSeq with the 250 base-pair paired-end protocol. Sequencing reads were merged into single reads and aligned to the human reference genome hg19 using BBMerge and BBMap functions of BBTools (Joint Genome Institute), obtaining an average of 78,864 aligned reads per experiment. SNVs and small indels in the regions corresponding to each crRNA were called using the CrispRVariants package in R (Lindsay et al. 2016) and the percentage of

edited alleles was estimated as the percentage of variant reads in each experiment.

Gene knockdown

Inducible gene knockdown was carried out using lentiviral pTRIPZ-TurboRFP (*MCM7*) or pSMART-TurboGFP (*PSMD3*) shRNA vectors (Dharmacon). For each gene, three shRNA vectors were tested (Table 4.2). Virus was produced by co-transfecting HEK293T cells with pTRIPZ or pSMART constructs alongside psPAX2 and pMD2.G vectors (Addgene) using Fugene HD (Promega). Viral supernatant was collected at 24 and 48 hours and used for two rounds of infection of OE19, OE33 or MFD1 cells, using 8μg/ml hexadimethrine bromide. Infected cells were selected after 48 hours with 2μg/ml puromycin for 7 days. To induce shRNA expression, cells were treated with 1μg/ml doxycycline for 16 hours. Gene expression with or without doxycycline was assessed by qRT-PCR using predesigned SYBR green primers (Sigma-Aldrich; Table 4.2). Cells with the highest level of knockdown were then sorted by FACS to isolate medium expressing cells (the middle 30% of cells based on TurboRFP or TurboGFP fluorescence). Gene expression after sorting was measured by qRT-PCR 24 hours post-induction with 0-1μg/ml doxycycline, to determine the concentration of doxycycline required to reduce expression to levels equivalent to FLO-1 cells. The determined concentrations of doxycycline used for proliferation assays were 0.05μg/ml for OE19 PSMD3 shRNA3, 0.25μg/ml for OE33 PSMD3 shRNA3, 0.25μg/ml for MFD1 PSMD3 shRNA3, 0.75μg/ml for MFD1 MCM7 shRNA3.

Cell proliferation

Cell proliferation was assessed using either the CellTiter 96 Non-Radioactive Cell Proliferation Assay or the CellTiter-Glo Luminescent Cell Viability Assay, according to manufacturer's instructions. Proliferation was assessed every 24 hours for three days, starting three hours after seeding the cells (time zero) to allow cell adhesion.

Briefly, $4.5x10^3$ cells/well were seeded on 96-well flat bottom plates in a final volume of 100ul per well. For the CellTiter 96 Non-Radioactive Cell Proliferation Assay, 15 ul of the Dye Solution was added into each well and cells were incubated at 37°C for two hours. The converted dye was released from the cells using 100µl of the Solubilisation Solution/Stop mix and absorbance was measured at 570nm after one hour using the Paradigm detection platform (Beckman Coulter). For the CellTiter-Glo Luminescent Cell Viability Assay, 100ul of the CellTiter-Glo reagent was added into each well and luminescence was measured after 30 minutes using the Paradigm detection platform (Beckman Coulter). Four replicates per condition were measured at each time point and each measure was normalised to the average time zero measure for each condition.

Flow cytometry

EdU incorporation and MCM loading were assessed using a modified version of the protocol described previously (Galanos et al. 2016). Briefly, in each condition, $3x10^6$ cells were pulsed for 30 minutes with 10µM EdU before washing in 1% BSA/PBS. Chromatin fractionation was performed by incubating on ice for 10 minutes in CSK buffer. Cells were then fixed in 4% formaldehyde/PBS for 10 minutes at room temperature before washing in 1%

BSA/PBS. Cells were permeabilised and barcoded (Krutzik and Nolan 2006) by incubating in 70% ethanol containing 0-15µg/ml Alexa Fluor 488 for 15 minutes, then washed twice in 1% BSA/PBS. Barcoded cells were subsequently pooled before incubating in primary antibody (mouse monoclonal anti-MCM7or rabbit polyclonal anti-MCM3) diluted 1:100 in 1% BSA/PBS for 1 hour. After washing in 1% BSA/PBS, samples were incubated for 30 minutes in secondary antibody (Alexa Fluor 555-conjugated donkey anti-mouse or donkey anti-rabbit) diluted 1:500 in 1% BSA/PBS, then washed again in 1% BSA/PBS. EdU labelling with Alexa Fluor 647 azide was performed using Click-iT EdU flow cytometry assay kit following the manufacturer's instructions before washing samples in 1% BSA/PBS. Samples were then incubated in 1% BSA/PBS containing RNase and 10mg/ml DAPI for 15 minutes before flow cytometry acquisition.

Flow cytometry acquisition and analysis

A 4-laser BD LSRII Fortessa flow cytometer (Beckton Dickinson, BD) was used for all flow cytometry studies. Lasers and filters used include: 407nm laser with 450/50 bandpass filter; 488nm laser with 505 longpass and 530/30 bandpass filters; 561nm laser with 570 longpass and 590/30 bandpass filters; 640nm laser with 670/14 bandpass filter. All samples were acquired using the FACS Diva programme (BD) and analysed using FlowJo software.

FlowJo 10.3 software (Treestar Inc, Oregon, USA) was used to analyse MCM loading onto chromatin and EdU incorporation. Compensation was performed manually with single colour controls, using BD FACSDiva software (BD Biosciences). Cells were gated to remove debris using FSC-A/SSC-A, then gated to isolate singlets using DAPI-H/DAPI-A. The cells were then separated by gating the barcoded populations using 488-A/DAPI-A. Cells were finally

separated into cell cycle gates (G1, S1-4, G2) based on EdU-647-A and DAPI-A, and the geometric mean fluorescence intensity was obtained for each channel (MCM-555 or EdU-647).

Chapter 7. Appendix

7.1 Known cancer genes with damaging alterations

List of 476 known cancer genes from the cancer gene census (Forbes et al. 2017) that acquire potential damaging alterations in 261 OACs. For each gene, the number of samples with damaging alterations in the 261 OACs and the number of samples in which the gene was considered as a driver is reported. This was based on the manual assessment of (a) the literature support for the cancer driver role in OAC, (b) the role as a tumour suppressor or oncogene as reported in (8) and the type of alterations found in OACs. For the 202 genes considered as drivers in at least one sample, the corresponding role is also reported.

Gene symbol	Description	Samples where the gene is altered (n)	Samples where the gene is driver (n)	Role in cancer
ABI1	abl-interactor 1	5	0	-
ABL1	v-abl Abelson murine leukemia viral O homolog 1	4	4	Oncogene
ABL2	c-abl O 2, non-receptor tyrosine kinase	3	0	-
ACKR3	atypical chemokine receptor 3	1	0	-
ACSL3	acyl-CoA synthetase long-chain family member 3	4	0	-
ACSL6	acyl-CoA synthetase long-chain family member 6	4	0	-
AFF1	AF4/FMR2 family, member 1	6	0	-
AFF3	AF4/FMR2 family, member 3	15	0	-
AFF4	AF4/FMR2 family, member 4	1	0	-
AKAP9	A kinase (PRKA) anchor protein (yotiao) 9	38	37	Oncogene
AKT1	v-akt murine thymoma viral O homolog 1	1	1	Oncogene
AKT2	v-akt murine thymoma viral O homolog 2	12	12	Oncogene
ALDH2	aldehyde dehydrogenase 2 family (mitochondrial)	3	0	-
ALK	anaplastic lymphoma kinase (Ki-1)	9	0	Oncogene
AMER1	APC membrane recruitment protein 1	1	0	Tumour suppressor
APC	adenomatous polyposis of the colon gene	23	20	Tumour suppressor
ARHGAP 26	Rho GTPase activating protein 26	5	5	Tumour suppressor
ARHGEF1 2	RHO guanine nucleotide exchange factor (GEF) 12 (LARG)	8	0	-
ARID1A	AT rich interactive domain 1A (SWI-like)	38	38	Tumour suppressor

ARID2	AT rich interactive domain 2	11	7	Tumour suppressor
ARNT	aryl hydrocarbon receptor nuclear translocator	11	0	-
ASPSCR1	alveolar soft part sarcoma chromosome region, candidate 1	1	0	-
ASXL1	additional sex combs like 1	33	1	Tumour suppressor
ATF1	activating transcription factor 1	3	0	-
ATIC	5-aminoimidazole-4-carboxamide ribonucleotide formyltransferase/IMP cyclohydrolase	3	0	-
ATM	ataxia telangiectasia mutated	16	13	Tumour suppressor
ATP1A1	ATPase, Na+/K+ transporting, alpha 1 polypeptide	5	4	Oncogene or Tumour suppressor
ATP2B3	ATPase, Ca++ transporting, plasma membrane 3	6	0	-
ATRX	alpha thalassemia/mental retardation syndrome X-linked	6	4	Tumour suppressor
AXIN1	axin 1	4	2	Tumour suppressor
BAP1	BRCA1 associated protein-1 (ubiquitin carboxy-terminal hydrolase)	3	3	Tumour suppressor
BCL11A	B-cell CLL/lymphoma 11A	11	0	-
BCL11B	B-cell CLL/lymphoma 11B (CTIP2)	1	1	Oncogene or Tumour suppressor
BCL3	B-cell CLL/lymphoma 3	18	0	-
BCL6	B-cell CLL/lymphoma 6	5	5	Oncogene
BCL7A	B-cell CLL/lymphoma 7A	4	0	-
BCL9	B-cell CLL/lymphoma 9	5	0	-
BCOR	BCL6 corepressor	2	2	Tumour suppressor
BCR	breakpoint cluster region	4	4	Oncogene
BIRC3	baculoviral IAP repeat-containing 3	9	9	Oncogene or Tumour suppressor
BLM	Bloom Syndrome	11	0	-
BMPR1A	bone morphogenetic protein receptor, type IA	5	0	-
BRAF	v-raf murine sarcoma viral O homolog B1	8	8	Oncogene
BRCA1	familial breast/ovarian cancer gene 1	8	0	Tumour suppressor
BRCA2	familial breast/ovarian cancer gene 2	23	0	Tumour suppressor
BRD3	bromodomain containing 3	5	4	Oncogene
BRD4	bromodomain containing 4	6	0	-
BRIP1	BRCA1 interacting protein C-terminal helicase 1	6	0	Tumour suppressor
BTG1	B-cell translocation gene 1,	1	0	Tumour

	anti-proliferative			suppressor
BUB1B	BUB1 budding uninhibited by benzimidazoles 1 homolog beta (yeast)	1	0	-
C15orf65	chromosome 15 open reading frame 65	1	0	-
CACNA1D	calcium channel, voltage-dependent, L type, alpha 1D subunit	6	5	Oncogene
CALR	calreticulin	3	1	Oncogene
CAMTA1	calmodulin binding transcription activator 1	16	14	Tumour suppressor
CANT1	calcium activated nucleotidase 1	3	0	-
CARD11	caspase recruitment domain family, member 11	6	5	Oncogene
CARS	cysteinyl-tRNA synthetase	2	0	-
CASC5	cancer susceptibility candidate 5	1	0	-
CASP8	caspase 8, apoptosis-related cysteine peptidase	6	2	Tumour suppressor
CBFA2T3	core-binding factor, runt domain, alpha subunit 2; translocated to, 3 (MTG-16)	4	0	-
CBFB	core-binding factor, beta subunit	1	1	Tumour suppressor
CBL	Cas-Br-M (murine) ecotropic retroviral transforming	11	10	Oncogene
CBLB	Cas-Br-M (murine) ecotropic retroviral transforming sequence b	3	2	Tumour suppressor
CBLC	Cas-Br-M (murine) ecotropic retroviral transforming sequence c	17	0	Oncogene or Tumour suppressor
CCDC6	coiled-coil domain containing 6	2	0	-
CCNB1IP1	cyclin B1 interacting protein 1, E3 ubiquitin protein ligase	6	0	-
CCND1	cyclin D1	34	34	Oncogene
CCND2	cyclin D2	7	6	Oncogene
CCND3	cyclin D3	25	25	Oncogene
CCNE1	cyclin E1	31	31	Oncogene
CD274	CD274 molecule	3	0	-
CD74	CD74 molecule, major histocompatibility complex, class II invariant chain	3	0	-
CD79A	CD79a molecule, immunoglobulin-associated alpha	8	8	Oncogene
CD79B	CD79b molecule, immunoglobulin-associated beta	3	2	Oncogene
CDC73	cell division cycle 73	4	1	Tumour suppressor
CDH1	cadherin 1, type 1, E-cadherin (epithelial) (ECAD)	8	4	Tumour suppressor

CDH11	cadherin 11, type 2, OB-cadherin (osteoblast)	12	0	-
CDK12	cyclin-dependent kinase 12	42	5	Tumour suppressor
CDK4	cyclin-dependent kinase 4	2	2	Oncogene
CDK6	cyclin-dependent kinase 6	40	40	Oncogene
CDKN2A	cyclin-dependent kinase inhibitor 2A (p16(INK4a)) gene	78	74	Tumour suppressor
CDKN2C	cyclin-dependent kinase inhibitor 2C (p18, inhibits CDK4)	2	1	Tumour suppressor
CDX2	caudal type homeo box transcription factor 2	13	0	-
CEBPA	CCAAT/enhancer binding protein (C/EBP), alpha	12	1	Tumour suppressor
CHCHD7	coiled-coil-helix-coiled-coil-helix domain containing 7	5	0	-
CHEK2	CHK2 checkpoint homolog (S. pombe)	4	2	Tumour suppressor
CHIC2	cysteine-rich hydrophobic domain 2	1	0	-
CHN1	-	5	0	-
CIC	capicua homolog	10	3	Tumour suppressor
CIITA	class II, major histocompatibility complex, transactivator	5	1	Tumour suppressor
CLP1	cleavage and polyadenylation factor I subunit 1	1	0	-
CLTC	clathrin, heavy polypeptide (Hc)	8	0	-
CLTCL1	clathrin, heavy polypeptide-like 1	4	0	-
CNBP	CCHC-type zinc finger, nucleic acid binding protein	1	0	-
CNOT3	CCR4-NOT transcription complex subunit 3	11	1	Tumour suppressor
CNTRL	centriolin	6	0	-
COL1A1	collagen, type I, alpha 1	7	0	-
COX6C	cytochrome c oxidase subunit VIc	14	0	-
CREB1	cAMP responsive element binding protein 1	1	0	-
CREB3L1	cAMP responsive element binding protein 3-like 1	9	0	-
CREB3L2	cAMP responsive element binding protein 3-like 2	1	0	-
CREBBP	CREB binding protein (CBP)	8	6	Tumour suppressor
CRLF2	cytokine receptor-like factor 2	23	0	Oncogene
CRTC1	CREB regulated transcription coactivator 1	4	4	Oncogene
CRTC3	CREB regulated transcription coactivator 3	9	0	-
CSF3R	colony stimulating factor 3 receptor (granulocyte)	4	4	Oncogene

CTNNB1	catenin (cadherin-associated protein), beta 1	7	7	Oncogene
CYLD	familial cylindromatosis gene	2	2	Tumour suppressor
DAXX	death-domain associated protein	8	0	Tumour suppressor
DDB2	damage-specific DNA binding protein 2	9	0	-
DDIT3	DNA-damage-inducible transcript 3	3	0	-
DDX10	DEAD (Asp-Glu-Ala-Asp) box polypeptide 10	10	7	Tumour suppressor
DDX5	DEAD (Asp-Glu-Ala-Asp) box polypeptide 5	2	0	-
DDX6	DEAD (Asp-Glu-Ala-Asp) box polypeptide 6	6	0	-
DEK	DEK O (DNA binding)	6	0	-
DICER1	dicer 1, ribonuclease type III	6	5	Tumour suppressor
DNM2	dynamin 2	9	1	Tumour suppressor
DNMT3A	DNA (cytosine-5-)-methyltransferase 3 alpha	1	0	Oncogene or Tumour suppressor
DUX4L1	double homeobox 4 like 1	1	0	-
EBF1	early B-cell factor 1	9	9	Oncogene
ECT2L	epithelial cell transforming sequence 2 O-like	6	0	Tumour suppressor
EGFR	epidermal growth factor receptor (erythroblastic leukemia viral (v-erb-b) O homolog, avian)	34	30	Oncogene
EIF4A2	eukaryotic translation initiation factor 4A, isoform 2	6	0	-
ELK4	ELK4, ETS-domain protein (SRF accessory protein 1)	3	0	-
ELL	ELL gene (11-19 lysine-rich leukemia gene)	4	0	-
ELN	elastin	12	12	Oncogene
EML4	echinoderm microtubule associated protein like 4	4	0	-
EP300	300 kd E1A-Binding protein gene	9	5	Tumour suppressor
EPS15	epidermal growth factor receptor pathway substrate 15 (AF1p)	3	0	-
ERBB2	v-erb-b2 erythroblastic leukemia viral O homolog 2, neuro/glioblastoma derived O homolog (avian)	51	48	Oncogene
ERC1	ELKS/RAB6-interacting/CAST family member 1	12	0	-
ERCC2	excision repair cross-complementing rodent repair deficiency, complementation group 2 (xeroderma pigmentosum D)	12	1	Tumour suppressor
ERCC3	excision repair cross-	2	0	-

	complementing rodent repair deficiency, complementation group 3 (xeroderma pigmentosum group B complementing)			
ERCC4	excision repair cross-complementing rodent repair deficiency, complementation group 4	4	0	-
ERCC5	excision repair cross-complementing rodent repair deficiency, complementation group 5 (xeroderma pigmentosum, complementation group G (Cockayne syndrome))	6	0	-
ERG	v-ets erythroblastosis virus E26 O like (avian)	3	2	Oncogene
ETV1	ets variant gene 1	9	0	-
ETV4	ets variant gene 4 (E1A enhancer binding protein, E1AF)	10	0	-
ETV5	ets variant gene 5	6	0	-
ETV6	ets variant gene 6 (TEL O)	12	12	Oncogene
EWSR1	Ewing sarcoma breakpoint region 1 (EWS)	1	0	-
EXT1	multiple exostoses type 1 gene	26	0	-
EXT2	multiple exostoses type 2 gene	6	1	Tumour suppressor
EZH2	enhancer of zeste homolog 2	6	5	Oncogene
EZR	ezrin	1	0	-
FAM46C	family with sequence similarity 46, member C	2	1	Tumour suppressor
FANCA	Fanconi anemia, complementation group A	3	1	Tumour suppressor
FANCC	Fanconi anemia, complementation group C	3	0	-
FANCD2	Fanconi anemia, complementation group D2	7	0	-
FANCE	Fanconi anemia, complementation group E	2	0	-
FANCF	Fanconi anemia, complementation group F	2	0	-
FANCG	Fanconi anemia, complementation group G	12	0	-
FAS	Fas cell surface death receptor	4	3	Tumour suppressor
FBXO11	F-box protein 11	5	2	Tumour suppressor
FBXW7	F-box and WD-40 domain protein 7 (archipelago homolog, Drosophila)	10	7	Tumour suppressor
FCGR2B	Fc fragment of IgG, low affinity IIb, receptor for (CD32)	5	0	-
FCRL4	Fc receptor-like 4	2	0	-
FEV	FEV protein - (HSRNAFEV)	1	0	-

FGFR1	fibroblast growth factor receptor 1	12	12	Oncogene
FGFR1OP	FGFR1 O partner (FOP)	2	0	-
FGFR2	fibroblast growth factor receptor 2	11	8	Oncogene
FGFR3	fibroblast growth factor receptor 3	4	3	Oncogene
FH	fumarate hydratase	5	1	Tumour suppressor
FHIT	fragile histidine triad gene	77	77	-
FIP1L1	FIP1 like 1 (S. cerevisiae)	9	0	-
FLCN	folliculin	4	0	-
FLI1	Friend leukemia virus integration 1	7	0	-
FLT3	fms-related tyrosine kinase 3	12	11	Oncogene
FNBP1	formin binding protein 1 (FBP17)	4	0	-
FOXA1	forkhead box A1	4	0	Oncogene
FOXL2	forkhead box L2	2	0	Oncogene
FOXO1	forkhead box O1	23	23	Oncogene or Tumour suppressor
FOXO4	forkhead box O4	1	0	-
FOXP1	forkhead box P1	9	0	-
FSTL3	follistatin-like 3 (secreted glycoprotein)	3	0	-
FUBP1	far upstream element (FUSE) binding protein 1	1	1	Tumour suppressor
FUS	fusion, derived from t(12;16) malignant liposarcoma	7	0	-
GAS7	growth arrest-specific 7	5	0	-
GATA2	GATA binding protein 2	1	1	Oncogene
GATA3	GATA binding protein 3	7	1	Tumour suppressor
GMPS	guanine monphosphate synthetase	3	0	-
GNA11	guanine nucleotide binding protein (G protein), alpha 11 (Gq class)	2	0	Oncogene
GNAQ	guanine nucleotide binding protein (G protein), q polypeptide	1	0	Oncogene
GNAS	guanine nucleotide binding protein (G protein), alpha stimulating activity polypeptide 1	33	30	Oncogene
GOLGA5	golgi autoantigen, golgin subfamily a, 5 (PTC5)	2	0	-
GOPC	golgi associated PDZ and coiled-coil motif containing	5	5	Oncogene
GPC3	glypican 3	4	0	-
GPHN	gephyrin (GPH)	8	0	-
H3F3A	H3 histone, family 3A	2	2	Oncogene
HERPUD1	homocysteine-inducible, endoplasmic reticulum stress-inducible, ubiquitin-	1	0	-

	like domain member 1			
HEY1	hairy/enhancer-of-split related with YRPW motif 1	10	0	-
HIP1	huntingtin interacting protein 1	13	0	-
HIST1H3B	histone cluster 1, H3b	7	0	Oncogene
HIST1H4I	histone 1, H4i (H4FM)	8	0	-
HLF	hepatic leukemia factor	5	0	-
HMGA1	high mobility group AT-hook 1	4	0	-
HMGA2	high mobility group AT-hook 2 (HMGIC)	11	0	-
HNF1A	HNF1 homeobox A	5	0	Tumour suppressor
HNRNPA2B1	heterogeneous nuclear ribonucleoprotein A2/B1	5	5	Oncogene
HOOK3	hook homolog 3	13	0	-
HOXA11	homeo box A11	7	0	-
HOXA13	homeo box A13	6	0	-
HOXA9	homeo box A9	6	0	-
HOXC11	homeo box C11	2	0	-
HOXC13	homeo box C13	1	0	-
HOXD11	homeo box D11	1	0	-
HOXD13	homeo box D13	2	0	-
HRAS	v-Ha-ras Harvey rat sarcoma viral O homolog	6	6	Oncogene
HSP90AA1	heat shock protein 90kDa alpha (cytosolic), class A member 1	7	0	-
HSP90AB1	heat shock protein 90kDa alpha (cytosolic), class B member 1	16	0	-
IDH1	isocitrate dehydrogenase 1 (NADP+), soluble	2	2	Oncogene
IDH2	isocitrate dehydrogenase 2 (NADP+), mitochondrial	7	7	Oncogene
IKZF1	IKAROS family zinc finger 1	12	4	Tumour suppressor
IL2	interleukin 2	1	0	-
IL21R	interleukin 21 receptor	10	0	-
IL6ST	interleukin 6 signal transducer (gp130, oncostatin M receptor)	3	0	-
IL7R	interleukin 7 receptor	15	12	Oncogene
IRF4	interferon regulatory factor 4	3	3	Oncogene or Tumour suppressor
ITK	IL2-inducible T-cell kinase	3	2	Oncogene
JAK1	Janus kinase 1	6	5	Oncogene
JAK2	Janus kinase 2	2	1	Oncogene
JAK3	Janus kinase 3	3	3	Oncogene
JAZF1	juxtaposed with another zinc finger gene 1	9	0	-

KAT6A	K(lysine) acetyltransferase 6A	13	12	Oncogene
KAT6B	K(lysine) acetyltransferase 6B	11	0	-
KCNJ5	potassium inwardly-rectifying channel; subfamily J; member 5	5	4	Oncogene
KDM5A	lysine (K)-specific demethylase 5A, JARID1A	11	0	-
KDM5C	lysine (K)-specific demethylase 5C (JARID1C)	1	1	Tumour suppressor
KDM6A	lysine (K)-specific demethylase 6A, UTX	9	4	Tumour suppressor
KIAA1549	KIAA1549	6	0	-
KIF5B	kinesin family member 5B	8	0	-
KIT	v-kit Hardy-Zuckerman 4 feline sarcoma viral O homolog	3	3	Oncogene
KLF4	Kruppel-like factor 4	2	2	Oncogene
KLF6	Kruppel-like factor 6	6	1	Tumour suppressor
KLK2	kallikrein-related peptidase 2	5	0	-
KMT2A	lysine (K)-specific methyltransferase 2A	12	11	Oncogene
KMT2C	lysine (K)-specific methyltransferase 2C	12	9	Tumour suppressor
KMT2D	lysine (K)-specific methyltransferase 2D	6	5	Tumour suppressor
KRAS	v-Ki-ras2 Kirsten rat sarcoma 2 viral O homolog	39	39	Oncogene
KTN1	kinectin 1 (kinesin receptor)	5	0	-
LASP1	LIM and SH3 protein 1	15	14	-
LCK	lymphocyte-specific protein tyrosine kinase	6	0	-
LCP1	lymphocyte cytosolic protein 1 (L-plastin)	19	0	-
LHFP	lipoma HMGIC fusion partner	20	15	-
LIFR	leukemia inhibitory factor receptor	19	0	-
LMO2	LIM domain only 2 (rhombotin-like 1) (RBTN2)	3	0	-
LPP	LIM domain containing preferred translocation partner in lipoma	10	0	-
LRIG3	leucine-rich repeats and immunoglobulin-like domains 3	6	0	-
LYL1	lymphoblastic leukemia derived sequence 1	1	0	-
MAF	v-maf musculoaponeurotic fibrosarcoma O homolog	1	0	-
MAFB	v-maf musculoaponeurotic fibrosarcoma O homolog B (avian)	30	0	-
MAML2	mastermind-like 2 (Drosophila)	11	11	Oncogene
MAP2K1	mitogen-activated protein kinase kinase 1	3	1	Oncogene

MAP2K4	mitogen-activated protein kinase kinase 4	8	7	Tumour suppressor
MAX	Myc associated factor X	4	0	Tumour suppressor
MDM2	Mdm2 p53 binding protein homolog	16	16	Oncogene
MDM4	Mdm4 p53 binding protein homolog	4	3	Oncogene
MECOM	MDS1 and EVI1 complex locus	21	21	Oncogene
MED12	mediator complex subunit 12	3	3	Oncogene
MEN1	multiple endocrine neoplasia type 1 gene	5	2	Tumour suppressor
MET	met proto-O (hepatocyte growth factor receptor)	13	12	Oncogene
MITF	microphthalmia-associated transcription factor	4	2	Oncogene
MKL1	megakaryoblastic leukemia (translocation) 1	8	0	-
MLF1	myeloid leukemia factor 1	2	0	-
MLH1	E.coli MutL homolog gene	3	0	Tumour suppressor
MLLT1	myeloid/lymphoid or mixed-lineage leukemia (trithorax homolog, Drosophila); translocated to, 1 (ENL)	1	0	-
MLLT10	myeloid/lymphoid or mixed-lineage leukemia (trithorax homolog, Drosophila); translocated to, 10 (AF10)	7	0	-
MLLT11	myeloid/lymphoid or mixed-lineage leukemia (trithorax homolog, Drosophila); translocated to, 11	11	0	-
MLLT3	myeloid/lymphoid or mixed-lineage leukemia (trithorax homolog, Drosophila); translocated to, 3 (AF9)	10	3	Oncogene
MLLT4	myeloid/lymphoid or mixed-lineage leukemia (trithorax homolog, Drosophila); translocated to, 4 (AF6)	8	0	-
MLLT6	myeloid/lymphoid or mixed-lineage leukemia (trithorax homolog, Drosophila); translocated to, 6 (AF17)	13	13	-
MN1	meningioma (disrupted in balanced translocation) 1	3	0	-
MNX1	motor neuron and pancreas homeobox 1	3	0	-
MPL	myeloproliferative leukaemia virus O, thrombopoietin receptor	5	0	Oncogene
MSH2	mutS homolog 2 (E. coli)	8	2	Tumour suppressor
MSI2	musashi homolog 2 (Drosophila)	10	0	-
MSN	moesin	5	0	-
MTCP1	mature T-cell proliferation 1	6	0	-

MUC1	mucin 1, transmembrane	13	0	-
MUTYH	mutY homolog (E. coli)	2	0	-
MYB	v-myb myeloblastosis viral O homolog	8	8	Oncogene
MYC	v-myc myelocytomatosis viral O homolog (avian)	55	55	Oncogene
MYCL	v-myc avian myelocytomatosis viral O lung carcinoma derived homolog	7	7	Oncogene
MYCN	v-myc myelocytomatosis viral related O, neuroblastoma derived (avian)	1	1	Oncogene
MYD88	myeloid differentiation primary response gene (88)	1	1	Oncogene
MYH11	myosin, heavy polypeptide 11, smooth muscle	10	0	-
MYH9	myosin, heavy polypeptide 9, non-muscle	7	0	-
NACA	nascent-polypeptide-associated complex alpha polypeptide	2	0	-
NBN	nibrin	15	0	-
NCOA1	nuclear receptor coactivator 1	6	0	-
NCOA2	nuclear receptor coactivator 2 (TIF2)	9	0	-
NDRG1	N-myc downstream regulated 1	23	0	-
NF1	neurofibromatosis type 1 gene	14	3	Tumour suppressor
NF2	neurofibromatosis type 2 gene	2	1	Tumour suppressor
NFE2L2	nuclear factor (erythroid-derived 2)-like 2 (NRF2)	4	1	Oncogene
NFIB	nuclear factor I/B	7	0	-
NFKB2	nuclear factor of kappa light polypeptide gene enhancer in B-cells 2 (p49/p100)	1	0	-
NIN	ninein (GSK3B interacting protein)	9	0	-
NKX2-1	NK2 homeobox 1	5	0	Oncogene
NONO	non-POU domain containing, octamer-binding	1	0	-
NOTCH1	Notch homolog 1, translocation-associated (Drosophila) (TAN1)	18	10	Tumour suppressor
NOTCH2	Notch homolog 2	6	0	Tumour suppressor
NPM1	nucleophosmin (nucleolar phosphoprotein B23, numatrin)	1	1	Tumour suppressor
NR4A3	nuclear receptor subfamily 4, group A, member 3 (NOR1)	2	0	-
NRAS	neuroblastoma RAS viral (v-ras) O homolog	5	3	Oncogene
NSD1	nuclear receptor binding SET domain protein 1	7	6	Oncogene

NT5C2	5'-nucleotidase, cytosolic II	3	2	Oncogene
NTRK1	neurotrophic tyrosine kinase, receptor, type 1	7	7	Oncogene
NTRK3	neurotrophic tyrosine kinase, receptor, type 3	11	10	Oncogene
NUMA1	nuclear mitotic apparatus protein 1	18	0	-
NUP214	nucleoporin 214kDa (CAN)	5	0	-
NUP98	nucleoporin 98kDa	5	0	-
NUTM1	NUT midline carcinoma, family member 1	2	0	-
NUTM2A	NUT family member 2A	1	0	-
NUTM2B	NUT family member 2B	7	0	-
OLIG2	oligodendrocyte lineage transcription factor 2 (BHLHB1)	4	0	-
P2RY8	purinergic receptor P2Y, G-protein coupled, 8	21	2	Oncogene
PAFAH1B2	platelet-activating factor acetylhydrolase, isoform Ib, beta subunit 30kDa	5	0	-
PALB2	partner and localizer of BRCA2	5	1	Tumour suppressor
PATZ1	-	1	0	-
PAX3	paired box gene 3	1	0	-
PAX5	paired box gene 5 (B-cell lineage specific activator protein)	12	0	Tumour suppressor
PAX7	paired box gene 7	3	0	-
PAX8	paired box gene 8	3	0	-
PBRM1	polybromo 1	12	8	Tumour suppressor
PBX1	pre-B-cell leukemia transcription factor 1	8	0	-
PCM1	pericentriolar material 1 (PTC4)	4	0	-
PCSK7	proprotein convertase subtilisin/kexin type 7	5	0	-
PDCD1LG2	programmed cell death 1 ligand 2	3	0	-
PDE4DIP	phosphodiesterase 4D interacting protein (myomegalin)	5	0	-
PDGFB	platelet-derived growth factor beta polypeptide (simian sarcoma viral (v-sis) O homolog)	2	0	-
PDGFRA	platelet-derived growth factor, alpha-receptor	6	6	Oncogene
PDGFRB	platelet-derived growth factor receptor, beta polypeptide	4	4	Oncogene
PER1	period homolog 1 (Drosophila)	2	0	-
PHF6	PHD finger protein 6	2	2	Tumour suppressor
PHOX2B	paired-like homeobox 2b	3	0	Tumour suppressor

PICALM	phosphatidylinositol binding clathrin assembly protein (CALM)	11	11	Oncogene
PIK3CA	phosphoinositide-3-kinase, catalytic, alpha polypeptide	23	23	Oncogene
PIK3R1	phosphoinositide-3-kinase, regulatory subunit 1 (alpha)	4	4	Tumour suppressor
PIM1	pim-1 O	8	8	Oncogene
PLAG1	pleiomorphic adenoma gene 1	5	0	-
PML	promyelocytic leukemia	1	0	-
PMS1	PMS1 postmeiotic segregation increased 1 (S. cerevisiae)	3	0	Tumour suppressor
PMS2	PMS2 postmeiotic segregation increased 2 (S. cerevisiae)	6	0	-
POT1	protection of telomeres 1	1	0	Oncogene
POU2AF1	POU domain, class 2, associating factor 1 (OBF1)	1	0	-
POU5F1	POU domain, class 5, transcription factor 1	18	0	-
PPARG	peroxisome proliferative activated receptor, gamma	2	0	-
PPP2R1A	protein phosphatase 2, regulatory subunit A, alpha	7	7	Oncogene
PRCC	papillary renal cell carcinoma (translocation-associated)	7	0	-
PRDM1	PR domain containing 1, with ZNF domain	1	0	Tumour suppressor
PRDM16	PR domain containing 16	9	0	-
PRF1	perforin 1 (pore forming protein)	1	0	-
PRKAR1A	protein kinase, cAMP-dependent, regulatory, type I, alpha (tissue specific extinguisher 1)	11	11	Oncogene or Tumour suppressor
PRRX1	paired related homeobox 1	6	0	-
PSIP1	PC4 and SFRS1 interacting protein 1 (LEDGF)	3	0	-
PTCH1	Homolog of Drosophila Patched gene	6	5	Tumour suppressor
PTEN	phosphatase and tensin homolog gene	11	10	Tumour suppressor
PTPN11	protein tyrosine phosphatase, non-receptor type 11	2	2	Oncogene
PTPRC	protein tyrosine phosphatase, receptor type, C	7	6	Tumour suppressor
RABEP1	rabaptin, RAB GTPase binding effector protein 1	7	0	-
RAC1	ras-related C3 botulinum toxin substrate 1 (rho family, small GTP binding protein Rac1)	4	2	Oncogene
RAD21	RAD21 homolog (S. pombe)	18	1	Tumour suppressor

RAD51B	RAD51 paralog B	14	0	-
RAF1	v-raf-1 murine leukemia viral O homolog 1	2	1	Oncogene
RALGDS	ral guanine nucleotide dissociation stimulator	4	0	-
RANBP17	RAN binding protein 17	6	0	-
RAP1GDS 1	RAP1, GTP-GDP dissociation stimulator 1	3	0	-
RARA	retinoic acid receptor, alpha	35	28	-
RB1	retinoblastoma gene	26	5	Tumour suppressor
RBM15	RNA binding motif protein 15	1	0	-
RECQL4	RecQ protein-like 4	19	19	-
REL	v-rel reticuloendotheliosis viral O homolog (avian)	2	1	Oncogene
RET	ret proto-O	8	8	Oncogene
RHOH	ras homolog family member H	1	0	-
RMI2	RecQ mediated genome instability 2	6	0	-
RNF213	ring finger protein 213	5	0	-
RNF43	ring finger protein 43	10	10	Tumour suppressor
ROS1	v-ros UR2 sarcoma virus O homolog 1 (avian)	7	0	Oncogene
RPL10	ribosomal protein L10	1	1	Oncogene
RPL22	ribosomal protein L22 (EAP)	4	4	Oncogene
RPN1	ribophorin I	2	0	-
RUNX1	runt-related transcription factor 1 (AML1)	32	32	Tumour suppressor
RUNX1T1	runt-related transcription factor 1; translocated to, 1 (cyclin D-related)	18	18	Oncogene
SBDS	Shwachman-Bodian-Diamond syndrome protein	12	10	-
SDC4	syndecan 4	35	0	-
SDHAF2	succinate dehydrogenase complex assembly factor 2	1	0	-
SDHB	succinate dehydrogenase complex, subunit B, iron sulfur (Ip)	2	0	-
SDHC	succinate dehydrogenase complex, subunit C, integral membrane protein, 15kDa	6	0	-
SDHD	succinate dehydrogenase complex, subunit D, integral membrane protein	1	0	-
SEPT5	septin 5	2	0	-
SEPT6	septin 6	1	0	-
SEPT9	septin 9	7	6	Oncogene
SET	SET translocation	3	0	-
SETBP1	SET binding protein 1	15	12	Oncogene
SETD2	SET domain containing 2	7	4	Tumour suppressor

SF3B1	splicing factor 3b, subunit 1, 155kDa	9	9	Oncogene
SFPQ	splicing factor proline/glutamine rich(polypyrimidine tract binding protein associated)	2	2	Oncogene
SH2B3	SH2B adaptor protein 3	2	0	Tumour suppressor
SH3GL1	SH3-domain GRB2-like 1 (EEN)	3	0	-
SLC45A3	solute carrier family 45, member 3	4	0	-
SMAD4	SMAD family member 4	49	45	Tumour suppressor
SMARCA 4	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily a, member 4	21	18	Tumour suppressor
SMARCB 1	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily b, member 1	5	3	Tumour suppressor
SMARCE 1	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily e, member 1	27	23	-
SMO	smoothened homolog (Drosophila)	3	2	Oncogene
SNX29	-	15	0	-
SOCS1	suppressor of cytokine signaling 1	6	0	Tumour suppressor
SOX2	SRY (sex determining region Y)-box 2	8	7	Oncogene
SPECC1	sperm antigen with calponin homology and coiled-coil domains 1	11	11	Oncogene
SRGAP3	SLIT-ROBO Rho GTPase activating protein 3	12	0	-
SRSF2	serine/arginine-rich splicing factor 2	3	0	Oncogene
SRSF3	serine/arginine-rich splicing factor 3	2	0	-
SS18	synovial sarcoma translocation, chromosome 18	11	8	-
SS18L1	synovial sarcoma translocation gene on chromosome 18-like 1	30	29	-
STAT3	signal transducer and activator of transcription 3 (acute-phase response factor)	9	8	Oncogene
STAT5B	signal transducer and activator of transcription 5B	8	8	Oncogene
STIL	SCL/TAL1 interrupting locus	2	0	-
STK11	serine/threonine kinase 11 gene (LKB1)	11	3	Tumour suppressor
SUFU	suppressor of fused homolog (Drosophila)	3	0	Tumour suppressor
SUZ12	suppressor of zeste 12	7	7	Oncogene

	homolog (Drosophila)			
TAF15	TAF15 RNA polymerase II, TATA box binding protein (TBP)-associated factor, 68kDa	3	1	-
TAL1	T-cell acute lymphocytic leukemia 1 (SCL)	1	0	-
TBL1XR1	transducin (beta)-like 1 X-linked receptor 1	7	1	Tumour suppressor
TCEA1	transcription elongation factor A (SII), 1	5	0	-
TCF12	transcription factor 12 (HTF4, helix-loop-helix transcription factors 4)	5	4	Oncogene
TCF3	transcription factor 3 (E2A immunoglobulin enhancer binding factors E12/E47)	4	0	-
TCF7L2	transcription factor 7-like 2	3	3	Oncogene
TERT	telomerase reverse transcriptase	17	0	Oncogene
TET1	tet methylcytosine dioxygenase 1	4	4	Oncogene or Tumour suppressor
TET2	tet O family member 2	4	3	Tumour suppressor
TFEB	transcription factor EB	25	0	-
TFG	TRK-fused gene	1	0	-
TFPT	TCF3 (E2A) fusion partner (in childhood leukaemia)	9	0	-
TFRC	transferrin receptor (p90, CD71)	5	0	-
THRAP3	thyroid hormone receptor associated protein 3 (TRAP150)	9	0	-
TMPRSS2	transmembrane protease, serine 2	7	4	Oncogene
TNFAIP3	tumor necrosis factor, alpha-induced protein 3	6	2	Tumour suppressor
TNFRSF14	tumor necrosis factor receptor superfamily, member 14 (herpesvirus entry mediator)	5	0	Tumour suppressor
TNFRSF17	tumor necrosis factor receptor superfamily, member 17	8	0	-
TOP1	topoisomerase (DNA) I	33	0	-
TP53	tumor protein p53	201	197	Tumour suppressor
TPM3	tropomyosin 3	11	0	-
TPM4	tropomyosin 4	7	0	-
TPR	translocated promoter region	6	0	-
TRAF7	tumour necrosis factor receptor-associated factor 7	3	0	Tumour suppressor
TRIM24	tripartite motif containing 24	3	0	-
TRIM27	tripartite motif-containing 27	6	0	-
TRIM33	tripartite motif-containing 33 (PTC7,TIF1G)	7	0	-

TRIP11	thyroid hormone receptor interactor 11	4	0	-
TRRAP	transformation/transcription domain-associated protein	39	36	Oncogene
TSC1	tuberous sclerosis 1 gene	3	1	Tumour suppressor
TSC2	tuberous sclerosis 2 gene	4	1	Tumour suppressor
TSHR	thyroid stimulating hormone receptor	3	0	Oncogene
TTL	tubulin tyrosine ligase	4	0	-
U2AF1	U2 small nuclear RNA auxiliary factor 1	2	2	Oncogene
UBR5	ubiquitin protein ligase E3 component n-recognin 5	20	3	Tumour suppressor
USP6	ubiquitin specific peptidase 6 (Tre-2 O)	2	0	-
VHL	von Hippel-Lindau syndrome gene	5	0	Tumour suppressor
VTI1A	vesicle transport through interaction with t-SNAREs homolog 1A	2	0	-
WHSC1	Wolf-Hirschhorn syndrome candidate 1(MMSET)	4	4	Oncogene
WHSC1L1	Wolf-Hirschhorn syndrome candidate 1-like 1 (NSD3)	15	0	-
WIF1	WNT inhibitory factor 1	10	0	-
WRN	Werner syndrome (RECQL2)	4	0	-
WT1	Wilms tumour 1 gene	3	0	Tumour suppressor
WWTR1	WW domain containing transcription regulator 1	2	0	-
XPA	xeroderma pigmentosum, complementation group A	1	0	-
XPC	xeroderma pigmentosum, complementation group C	4	0	-
XPO1	exportin 1 (CRM1 homolog, yeast)	1	1	Oncogene
YWHAE	tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, epsilon polypeptide (14-3-3 epsilon)	9	0	-
ZBTB16	zinc finger and BTB domain containing 16	4	4	Oncogene
ZMYM2	-	11	9	Oncogene or Tumour suppressor
ZNF331	zinc finger protein 331	11	0	-
ZNF384	zinc finger protein 384 (CIZ/NMP4)	9	0	-
ZNF521	zinc finger protein 521	36	35	Oncogene
ZRSR2	zinc finger (CCCH type), RNA-binding motif and serine/arginine rich 2	1	0	Tumour suppressor

7.2 Cancer helper genes in 261 OACs

List of 952 cancer helper genes in the 261 OACs. For each gene reported are the gene description; the number and percentage of samples where it is altered.

Gene	Description	Total altered samples (n)	Total altered samples (%)
ABCA1	ATP-binding cassette, sub-family A (ABC1), member 1	6	2.3
ABCC1	ATP-binding cassette, sub-family C (CFTR/MRP), member 1	10	3.8
ABI2	abl-interactor 2	2	0.8
ABLIM1	actin binding LIM protein 1	1	0.4
ABLIM3	actin binding LIM protein family, member 3	1	0.4
ABR	active BCR-related	4	1.5
ACACA	acetyl-CoA carboxylase alpha	7	2.7
ACLY	ATP citrate lyase	5	1.9
ACTB	actin, beta	2	0.8
ACTG1	actin, gamma 1	2	0.8
ACTN4	actinin, alpha 4	14	5.4
ACVR1	activin A receptor, type I	1	0.4
ACVR2B	activin A receptor, type IIB	1	0.4
ADAM15	ADAM metallopeptidase domain 15	1	0.4
ADAM22	ADAM metallopeptidase domain 22	2	0.8
ADAM9	ADAM metallopeptidase domain 9	4	1.5
ADARB1	adenosine deaminase, RNA-specific, B1	1	0.4
ADD2	adducin 2 (beta)	1	0.4
ADRM1	adhesion regulating molecule 1	4	1.5
AGO1	argonaute RISC catalytic component 1	3	1.1
AGO2	argonaute RISC catalytic component 2	16	6.1
AGPAT5	1-acylglycerol-3-phosphate O-acyltransferase 5	1	0.4
AGTR1	angiotensin II receptor, type 1	1	0.4
AHSA1	AHA1, activator of heat shock 90kDa protein ATPase homolog 1 (yeast)	1	0.4
AIP	aryl hydrocarbon receptor interacting protein	4	1.5
AJUBA	ajuba LIM protein	2	0.8
ALDH1B1	aldehyde dehydrogenase 1 family, member B1	1	0.4
ALDH9A1	aldehyde dehydrogenase 9 family, member A1	2	0.8
ALDOA	aldolase A, fructose-bisphosphate	1	0.4
AMPH	amphiphysin	2	0.8
ANK3	ankyrin 3, node of Ranvier (ankyrin G)	2	0.8
ANKRD17	ankyrin repeat domain 17	1	0.4
ANKRD28	ankyrin repeat domain 28	1	0.4
ANKRD29	ankyrin repeat domain 29	1	0.4

ANKRD52	ankyrin repeat domain 52	1	0.4
ANO6	anoctamin 6	1	0.4
ANP32E	acidic (leucine-rich) nuclear phosphoprotein 32 family, member E	1	0.4
ANXA1	annexin A1	1	0.4
ANXA11	annexin A11	7	2.7
AP1B1	adaptor-related protein complex 1, beta 1 subunit	1	0.4
AP2B1	adaptor-related protein complex 2, beta 1 subunit	1	0.4
APLP2	amyloid beta (A4) precursor-like protein 2	3	1.1
APP	amyloid beta (A4) precursor protein	4	1.5
AQP1	aquaporin 1 (Colton blood group)	2	0.8
ARCN1	archain 1	1	0.4
ARG1	arginase 1	1	0.4
ARHGAP1	Rho GTPase activating protein 1	6	2.3
ARHGAP12	Rho GTPase activating protein 12	1	0.4
ARHGAP17	Rho GTPase activating protein 17	4	1.5
ARHGAP29	Rho GTPase activating protein 29	1	0.4
ARHGDIA	Rho GDP dissociation inhibitor (GDI) alpha	1	0.4
ARHGEF10	Rho guanine nucleotide exchange factor (GEF) 10	1	0.4
ARHGEF11	Rho guanine nucleotide exchange factor (GEF) 11	1	0.4
ARHGEF2	Rho/Rac guanine nucleotide exchange factor (GEF) 2	9	3.4
ARID3A	AT rich interactive domain 3A (BRIGHT-like)	3	1.1
ARID4B	AT rich interactive domain 4B (RBP1-like)	3	1.1
ARNT2	aryl-hydrocarbon receptor nuclear translocator 2	1	0.4
ARPC2	actin related protein 2/3 complex, subunit 2, 34kDa	1	0.4
ASAP1	ArfGAP with SH3 domain, ankyrin repeat and PH domain 1	20	7.7
ASH1L	ash1 (absent, small, or homeotic)-like (Drosophila)	1	0.4
ATAD2	ATPase family, AAA domain containing 2	2	0.8
ATF3	activating transcription factor 3	1	0.4
ATF6	activating transcription factor 6	1	0.4
ATN1	atrophin 1	4	1.5
ATP10A	ATPase, class V, type 10A	1	0.4
ATP13A3	ATPase type 13A3	1	0.4
ATP2A2	ATPase, Ca++ transporting, cardiac muscle, slow twitch 2	2	0.8
ATP6V0A1	ATPase, H+ transporting, lysosomal V0 subunit a1	1	0.4
ATP6V1B2	ATPase, H+ transporting, lysosomal 56/58kDa, V1 subunit B2	1	0.4
ATP6V1C1	ATPase, H+ transporting, lysosomal 42kDa, V1 subunit C1	5	1.9
ATP7A	ATPase, Cu++ transporting, alpha polypeptide	2	0.8
ATP7B	ATPase, Cu++ transporting, beta polypeptide	12	4.6

ATP8B2	ATPase, aminophospholipid transporter, class I, type 8B, member 2	4	1.5
ATXN7	ataxin 7	2	0.8
AXIN2	axin 2	1	0.4
AXL	AXL receptor tyrosine kinase	8	3.1
B4GALT1	UDP-Gal:betaGlcNAc beta 1,4-galactosyltransferase, polypeptide 1	3	1.1
B4GALT3	UDP-Gal:betaGlcNAc beta 1,4-galactosyltransferase, polypeptide 3	1	0.4
B4GALT5	UDP-Gal:betaGlcNAc beta 1,4-galactosyltransferase, polypeptide 5	1	0.4
BACE1	beta-site APP-cleaving enzyme 1	3	1.1
BACH1	BTB and CNC homology 1, basic leucine zipper transcription factor 1	1	0.4
BAG6	BCL2-associated athanogene 6	24	9.2
BAZ1B	bromodomain adjacent to zinc finger domain, 1B	5	1.9
BCAT1	branched chain amino-acid transaminase 1, cytosolic	6	2.3
BDNF	brain-derived neurotrophic factor	1	0.4
BIRC2	baculoviral IAP repeat containing 2	1	0.4
BNIP3L	BCL2/adenovirus E1B 19kDa interacting protein 3-like	2	0.8
BPTF	bromodomain PHD finger transcription factor	3	1.1
BRD2	bromodomain containing 2	1	0.4
BRPF1	bromodomain and PHD finger containing, 1	1	0.4
BTG2	BTG family, member 2	2	0.8
BTG3	BTG family, member 3	2	0.8
BZW2	basic leucine zipper and W2 domains 2	1	0.4
C3orf62	chromosome 3 open reading frame 62	1	0.4
CA13	carbonic anhydrase XIII	4	1.5
CAB39L	calcium binding protein 39-like	1	0.4
CABLES1	Cdk5 and Abl enzyme substrate 1	11	4.2
CACNA1C	calcium channel, voltage-dependent, L type, alpha 1C subunit	1	0.4
CALM2	calmodulin 2 (phosphorylase kinase, delta)	2	0.8
CALU	calumenin	4	1.5
CAMK1D	calcium/calmodulin-dependent protein kinase ID	1	0.4
CAPN1	calpain 1, (mu/I) large subunit	1	0.4
CAPN15	calpain 15	1	0.4
CAPN2	calpain 2, (m/II) large subunit	1	0.4
CAPNS1	calpain, small subunit 1	6	2.3
CAPRIN1	cell cycle associated protein 1	1	0.4
CARD6	caspase recruitment domain family, member 6	1	0.4
CARHSP1	calcium regulated heat stable protein 1, 24kDa	2	0.8
CASP9	caspase 9, apoptosis-related cysteine peptidase	1	0.4
CAT	catalase	1	0.4
CBFA2T2	core-binding factor, runt domain, alpha subunit 2; translocated to, 2	1	0.4

CBS	cystathionine-beta-synthase	1	0.4
CCAR2	cell cycle and apoptosis regulator 2	2	0.8
CCNA1	cyclin A1	1	0.4
CCNA2	cyclin A2	1	0.4
CCNE2	cyclin E2	1	0.4
CCNG2	cyclin G2	1	0.4
CCNL1	cyclin L1	1	0.4
CCNY	cyclin Y	2	0.8
CCT3	chaperonin containing TCP1, subunit 3 (gamma)	1	0.4
CD151	CD151 molecule (Raph blood group)	1	0.4
CD2AP	CD2-associated protein	9	3.4
CD36	CD36 molecule (thrombospondin receptor)	2	0.8
CD44	CD44 molecule (Indian blood group)	2	0.8
CDC20	cell division cycle 20	1	0.4
CDC25A	cell division cycle 25A	1	0.4
CDC25B	cell division cycle 25B	7	2.7
CDC42	cell division cycle 42	1	0.4
CDC6	cell division cycle 6	4	1.5
CDCA7L	cell division cycle associated 7-like	3	1.1
CDH2	cadherin 2, type 1, N-cadherin (neuronal)	3	1.1
CDK14	cyclin-dependent kinase 14	7	2.7
CDK18	cyclin-dependent kinase 18	3	1.1
CDK19	cyclin-dependent kinase 19	2	0.8
CDK3	cyclin-dependent kinase 3	1	0.4
CDK5R1	cyclin-dependent kinase 5, regulatory subunit 1 (p35)	1	0.4
CDKN1A	cyclin-dependent kinase inhibitor 1A (p21, Cip1)	4	1.5
CDKN2B	cyclin-dependent kinase inhibitor 2B (p15, inhibits CDK4)	4	1.5
CDKN2D	cyclin-dependent kinase inhibitor 2D (p19, inhibits CDK4)	1	0.4
CEBPB	CCAAT/enhancer binding protein (C/EBP), beta	3	1.1
CELF2	CUGBP, Elav-like family member 2	1	0.4
CERS2	ceramide synthase 2	12	4.6
CHAMP1	chromosome alignment maintaining phosphoprotein 1	1	0.4
CHD4	chromodomain helicase DNA binding protein 4	2	0.8
CHD7	chromodomain helicase DNA binding protein 7	3	1.1
CHD8	chromodomain helicase DNA binding protein 8	2	0.8
CHRAC1	chromatin accessibility complex 1	1	0.4
CKAP5	cytoskeleton associated protein 5	3	1.1
CLIC1	chloride intracellular channel 1	3	1.1
CLIP1	CAP-GLY domain containing linker protein 1	1	0.4
CLK1	CDC-like kinase 1	1	0.4
CLN8	ceroid-lipofuscinosis, neuronal 8 (epilepsy,	1	0.4

	progressive with mental retardation)		
CNN3	calponin 3, acidic	1	0.4
CNOT1	CCR4-NOT transcription complex, subunit 1	5	1.9
CNOT7	CCR4-NOT transcription complex, subunit 7	1	0.4
CNTNAP2	contactin associated protein-like 2	1	0.4
COCH	cochlin	1	0.4
COL4A1	collagen, type IV, alpha 1	8	3.1
COL4A2	collagen, type IV, alpha 2	8	3.1
COL4A5	collagen, type IV, alpha 5	1	0.4
COPA	coatomer protein complex, subunit alpha	3	1.1
COPB1	coatomer protein complex, subunit beta 1	1	0.4
COPS3	COP9 signalosome subunit 3	1	0.4
CORO1A	coronin, actin binding protein, 1A	2	0.8
CORO1B	coronin, actin binding protein, 1B	1	0.4
CORO1C	coronin, actin binding protein, 1C	1	0.4
CPSF6	cleavage and polyadenylation specific factor 6, 68kDa	2	0.8
CPT1A	carnitine palmitoyltransferase 1A (liver)	8	3.1
CREM	cAMP responsive element modulator	1	0.4
CRIP2	cysteine-rich protein 2	1	0.4
CRKL	v-crk avian sarcoma virus CT10 oncogene homolog-like	3	1.1
CRTAP	cartilage associated protein	1	0.4
CRY2	cryptochrome circadian clock 2	1	0.4
CSDE1	cold shock domain containing E1, RNA-binding	3	1.1
CTCF	CCCTC-binding factor (zinc finger protein)	1	0.4
CTNND1	catenin (cadherin-associated protein), delta 1	2	0.8
CTPS1	CTP synthase 1	3	1.1
CTSB	cathepsin B	2	0.8
CTSD	cathepsin D	1	0.4
CUL3	cullin 3	1	0.4
CUX1	cut-like homeobox 1	9	3.4
CYP2E1	cytochrome P450, family 2, subfamily E, polypeptide 1	1	0.4
DAAM1	dishevelled associated activator of morphogenesis 1	1	0.4
DCAF7	DDB1 and CUL4 associated factor 7	3	1.1
DCLRE1B	DNA cross-link repair 1B	1	0.4
DDHD2	DDHD domain containing 2	4	1.5
DDR1	discoidin domain receptor tyrosine kinase 1	1	0.4
DDX17	DEAD (Asp-Glu-Ala-Asp) box helicase 17	2	0.8
DDX3Y	DEAD (Asp-Glu-Ala-Asp) box helicase 3, Y-linked	1	0.4
DEGS1	delta(4)-desaturase, sphingolipid 1	1	0.4
DENND4B	DENN/MADD domain containing 4B	1	0.4
DHCR7	7-dehydrocholesterol reductase	1	0.4

DHX29	DEAH (Asp-Glu-Ala-His) box polypeptide 29	1	0.4
DIP2A	DIP2 disco-interacting protein 2 homolog A (Drosophila)	2	0.8
DLC1	DLC1 Rho GTPase activating protein	24	9.2
DLG2	discs, large homolog 2 (Drosophila)	5	1.9
DMWD	dystrophia myotonica, WD repeat containing	2	0.8
DMXL2	Dmx-like 2	1	0.4
DNAJB5	DnaJ (Hsp40) homolog, subfamily B, member 5	10	3.8
DNAJC5	DnaJ (Hsp40) homolog, subfamily C, member 5	4	1.5
DNMBP	dynamin binding protein	1	0.4
DNMT1	DNA (cytosine-5-)-methyltransferase 1	4	1.5
DNMT3B	DNA (cytosine-5-)-methyltransferase 3 beta	25	9.6
DOCK4	dedicator of cytokinesis 4	1	0.4
DOCK5	dedicator of cytokinesis 5	2	0.8
DOCK8	dedicator of cytokinesis 8	2	0.8
DOPEY1	dopey family member 1	1	0.4
DPY19L4	dpy-19-like 4 (C. elegans)	1	0.4
DPYSL2	dihydropyrimidinase-like 2	2	0.8
DSP	desmoplakin	1	0.4
DST	dystonin	2	0.8
DSTN	destrin (actin depolymerizing factor)	1	0.4
DTNB	dystrobrevin, beta	1	0.4
DTX2	deltex 2, E3 ubiquitin ligase	1	0.4
DVL1	dishevelled segment polarity protein 1	1	0.4
DYNC1I1	dynein, cytoplasmic 1, intermediate chain 1	1	0.4
DYRK1A	dual-specificity tyrosine-(Y)-phosphorylation regulated kinase 1A	1	0.4
DYRK1B	dual-specificity tyrosine-(Y)-phosphorylation regulated kinase 1B	12	4.6
DYRK2	dual-specificity tyrosine-(Y)-phosphorylation regulated kinase 2	9	3.4
DYSF	dysferlin	2	0.8
E2F1	E2F transcription factor 1	29	11.1
E2F2	E2F transcription factor 2	1	0.4
E2F3	E2F transcription factor 3	1	0.4
E2F4	E2F transcription factor 4, p107/p130-binding	1	0.4
E2F5	E2F transcription factor 5, p130-binding	2	0.8
E2F6	E2F transcription factor 6	1	0.4
EAF1	ELL associated factor 1	1	0.4
ECT2	epithelial cell transforming 2	2	0.8
EDF1	endothelial differentiation-related factor 1	1	0.4
EDNRA	endothelin receptor type A	1	0.4
EEA1	early endosome antigen 1	1	0.4
EFEMP2	EGF containing fibulin-like extracellular matrix protein 2	2	0.8
EFHC2	EF-hand domain (C-terminal) containing 2	1	0.4

EFHD2	EF-hand domain family, member D2	1	0.4
EFNB2	ephrin-B2	2	0.8
EGLN3	egl-9 family hypoxia-inducible factor 3	5	1.9
EGR1	early growth response 1	1	0.4
EHMT1	euchromatic histone-lysine N-methyltransferase 1	1	0.4
EHMT2	euchromatic histone-lysine N-methyltransferase 2	11	4.2
EIF3C	eukaryotic translation initiation factor 3, subunit C	1	0.4
EIF4A1	eukaryotic translation initiation factor 4A1	2	0.8
EIF4E	eukaryotic translation initiation factor 4E	1	0.4
EIF4EBP1	eukaryotic translation initiation factor 4E binding protein 1	3	1.1
EIF4G1	eukaryotic translation initiation factor 4 gamma, 1	4	1.5
EIF4G2	eukaryotic translation initiation factor 4 gamma, 2	1	0.4
EIF4G3	eukaryotic translation initiation factor 4 gamma, 3	2	0.8
ELAVL1	ELAV like RNA binding protein 1	1	0.4
ELOVL1	ELOVL fatty acid elongase 1	1	0.4
ELOVL5	ELOVL fatty acid elongase 5	2	0.8
ENAH	enabled homolog (Drosophila)	3	1.1
ENPP4	ectonucleotide pyrophosphatase/phosphodiesterase 4 (putative)	2	0.8
EPB41	erythrocyte membrane protein band 4.1	2	0.8
EPB41L2	erythrocyte membrane protein band 4.1-like 2	1	0.4
EPB41L3	erythrocyte membrane protein band 4.1-like 3	1	0.4
EPC1	enhancer of polycomb homolog 1 (Drosophila)	2	0.8
EPHA2	EPH receptor A2	2	0.8
EPHA4	EPH receptor A4	4	1.5
EPHB1	EPH receptor B1	1	0.4
EPHB2	EPH receptor B2	3	1.1
EPHB6	EPH receptor B6	1	0.4
ERBB2IP	erbb2 interacting protein	2	0.8
ERLIN1	ER lipid raft associated 1	1	0.4
ERLIN2	ER lipid raft associated 2	6	2.3
ERO1L	ERO1-like (S. cerevisiae)	3	1.1
ESYT1	extended synaptotagmin-like protein 1	5	1.9
ETS1	v-ets avian erythroblastosis virus E26 oncogene homolog 1	4	1.5
ETS2	v-ets avian erythroblastosis virus E26 oncogene homolog 2	1	0.4
EXOC7	exocyst complex component 7	1	0.4
EXOSC8	exosome component 8	1	0.4
EZH1	enhancer of zeste 1 polycomb repressive complex 2 subunit	1	0.4
FAM126A	family with sequence similarity 126, member A	1	0.4

FAM126B	family with sequence similarity 126, member B	2	0.8
FAM208B	family with sequence similarity 208, member B	1	0.4
FAM49B	family with sequence similarity 49, member B	5	1.9
FAM57A	family with sequence similarity 57, member A	2	0.8
FANCI	Fanconi anemia, complementation group I	1	0.4
FAR1	fatty acyl CoA reductase 1	1	0.4
FASN	fatty acid synthase	1	0.4
FBXO28	F-box protein 28	1	0.4
FCHSD2	FCH and double SH3 domains 2	4	1.5
FDFT1	farnesyl-diphosphate farnesyltransferase 1	3	1.1
FER	fer (fps/fes related) tyrosine kinase	1	0.4
FES	FES proto-oncogene, tyrosine kinase	1	0.4
FGF2	fibroblast growth factor 2 (basic)	1	0.4
FHL3	four and a half LIM domains 3	1	0.4
FKBP10	FK506 binding protein 10, 65 kDa	3	1.1
FKBP4	FK506 binding protein 4, 59kDa	4	1.5
FKBP5	FK506 binding protein 5	1	0.4
FLNA	filamin A, alpha	3	1.1
FLNB	filamin B, beta	3	1.1
FLT1	fms-related tyrosine kinase 1	8	3.1
FMNL1	formin-like 1	2	0.8
FNDC3B	fibronectin type III domain containing 3B	5	1.9
FOS	FBJ murine osteosarcoma viral oncogene homolog	1	0.4
FOXP2	forkhead box P2	1	0.4
FRS2	fibroblast growth factor receptor substrate 2	1	0.4
FTH1	ferritin, heavy polypeptide 1	1	0.4
FUBP3	far upstream element (FUSE) binding protein 3	1	0.4
FURIN	furin (paired basic amino acid cleaving enzyme)	1	0.4
FXR2	fragile X mental retardation, autosomal homolog 2	1	0.4
FZD1	frizzled class receptor 1	1	0.4
G3BP2	GTPase activating protein (SH3 domain) binding protein 2	5	1.9
GAB1	GRB2-associated binding protein 1	1	0.4
GAB2	GRB2-associated binding protein 2	1	0.4
GABRA5	gamma-aminobutyric acid (GABA) A receptor, alpha 5	1	0.4
GABRB3	gamma-aminobutyric acid (GABA) A receptor, beta 3	1	0.4
GALNT2	polypeptide N-acetylgalactosaminyltransferase 2	1	0.4
GANAB	glucosidase, alpha; neutral AB	4	1.5
GATA6	GATA binding protein 6	9	3.4
GATAD2B	GATA zinc finger domain containing 2B	4	1.5
GBAS	glioblastoma amplified sequence	4	1.5
GCN1L1	GCN1 general control of amino-acid synthesis	2	0.8

	1-like 1 (yeast)		
GDAP1	ganglioside induced differentiation associated protein 1	1	0.4
GDI2	GDP dissociation inhibitor 2	2	0.8
GEM	GTP binding protein overexpressed in skeletal muscle	5	1.9
GEMIN5	gem (nuclear organelle) associated protein 5	1	0.4
GFOD1	glucose-fructose oxidoreductase domain containing 1	2	0.8
GGA3	golgi-associated, gamma adaptin ear containing, ARF binding protein 3	6	2.3
GIGYF1	GRB10 interacting GYF protein 1	8	3.1
GIGYF2	GRB10 interacting GYF protein 2	1	0.4
GLI1	GLI family zinc finger 1	2	0.8
GLI2	GLI family zinc finger 2	1	0.4
GLI3	GLI family zinc finger 3	5	1.9
GNA12	guanine nucleotide binding protein (G protein) alpha 12	5	1.9
GNA13	guanine nucleotide binding protein (G protein), alpha 13	2	0.8
GNAI2	guanine nucleotide binding protein (G protein), alpha inhibiting activity polypeptide 2	1	0.4
GNB1	guanine nucleotide binding protein (G protein), beta polypeptide 1	2	0.8
GNB2	guanine nucleotide binding protein (G protein), beta polypeptide 2	1	0.4
GNS	glucosamine (N-acetyl)-6-sulfatase	6	2.3
GOLT1B	golgi transport 1B	1	0.4
GPATCH8	G patch domain containing 8	1	0.4
GPM6B	glycoprotein M6B	1	0.4
GRB10	growth factor receptor-bound protein 10	1	0.4
GRB2	growth factor receptor-bound protein 2	3	1.1
GRIA1	glutamate receptor, ionotropic, AMPA 1	1	0.4
GRINA	glutamate receptor, ionotropic, N-methyl D-aspartate-associated protein 1 (glutamate binding)	5	1.9
GRSF1	G-rich RNA sequence binding factor 1	2	0.8
GSN	gelsolin	1	0.4
H2AFY	H2A histone family, member Y	1	0.4
HCFC1	host cell factor C1	2	0.8
HCK	HCK proto-oncogene, Src family tyrosine kinase	3	1.1
HDAC7	histone deacetylase 7	1	0.4
HDLBP	high density lipoprotein binding protein	2	0.8
HEATR2	HEAT repeat containing 2	5	1.9
HES1	hes family bHLH transcription factor 1	3	1.1
HEXA	hexosaminidase A (alpha polypeptide)	1	0.4
HGF	hepatocyte growth factor (hepapoietin A; scatter factor)	5	1.9
HIP1R	huntingtin interacting protein 1 related	1	0.4
HIPK1	homeodomain interacting protein kinase 1	2	0.8

HIPK2	homeodomain interacting protein kinase 2	6	2.3
HIPK3	homeodomain interacting protein kinase 3	5	1.9
HIST2H2AC	histone cluster 2, H2ac	1	0.4
HIST2H2BE	histone cluster 2, H2be	1	0.4
HIVEP1	human immunodeficiency virus type I enhancer binding protein 1	5	1.9
HMGB1	high mobility group box 1	5	1.9
HMOX1	heme oxygenase (decycling) 1	1	0.4
HNRNPDL	heterogeneous nuclear ribonucleoprotein D-like	1	0.4
HNRNPF	heterogeneous nuclear ribonucleoprotein F	1	0.4
HNRNPL	heterogeneous nuclear ribonucleoprotein L	3	1.1
HNRNPM	heterogeneous nuclear ribonucleoprotein M	2	0.8
HNRNPUL1	heterogeneous nuclear ribonucleoprotein U-like 1	10	3.8
HSPA4	heat shock 70kDa protein 4	1	0.4
HSPH1	heat shock 105kDa/110kDa protein 1	14	5.4
HUS1	HUS1 checkpoint homolog (S. pombe)	1	0.4
IARS	isoleucyl-tRNA synthetase	2	0.8
ICK	intestinal cell (MAK-like) kinase	2	0.8
ID3	inhibitor of DNA binding 3, dominant negative helix-loop-helix protein	1	0.4
IGF1R	insulin-like growth factor 1 receptor	4	1.5
IGFBP3	insulin-like growth factor binding protein 3	3	1.1
IGFBP4	insulin-like growth factor binding protein 4	1	0.4
IKBKAP	inhibitor of kappa light polypeptide gene enhancer in B-cells, kinase complex-associated protein	1	0.4
IKBKB	inhibitor of kappa light polypeptide gene enhancer in B-cells, kinase beta	4	1.5
IKZF3	IKAROS family zinc finger 3 (Aiolos)	2	0.8
IL1RAP	interleukin 1 receptor accessory protein	1	0.4
IMPDH2	IMP (inosine 5-monophosphate) dehydrogenase 2	1	0.4
ING1	inhibitor of growth family, member 1	2	0.8
INPP5D	inositol polyphosphate-5-phosphatase, 145kDa	1	0.4
INPPL1	inositol polyphosphate phosphatase-like 1	3	1.1
IPO5	importin 5	1	0.4
IQGAP1	IQ motif containing GTPase activating protein 1	1	0.4
IRAK1	interleukin-1 receptor-associated kinase 1	5	1.9
IRAK2	interleukin-1 receptor-associated kinase 2	1	0.4
IRAK3	interleukin-1 receptor-associated kinase 3	1	0.4
IRF2BP2	interferon regulatory factor 2 binding protein 2	1	0.4
IRF7	interferon regulatory factor 7	1	0.4
IRS2	insulin receptor substrate 2	5	1.9
ITGA2	integrin, alpha 2 (CD49B, alpha 2 subunit of VLA-2 receptor)	2	0.8
ITGA3	integrin, alpha 3 (antigen CD49C, alpha 3 subunit of VLA-3 receptor)	1	0.4

ITGA5	integrin, alpha 5 (fibronectin receptor, alpha polypeptide)	4	1.5
ITGA6	integrin, alpha 6	3	1.1
ITGAV	integrin, alpha V	2	0.8
ITGB1	integrin, beta 1 (fibronectin receptor, beta polypeptide, antigen CD29 includes MDF2, MSK12)	5	1.9
ITM2B	integral membrane protein 2B	1	0.4
ITPR1	inositol 1,4,5-trisphosphate receptor, type 1	4	1.5
ITPR3	inositol 1,4,5-trisphosphate receptor, type 3	1	0.4
JAG1	jagged 1	1	0.4
JUNB	jun B proto-oncogene	1	0.4
JUP	junction plakoglobin	14	5.4
KANK2	KN motif and ankyrin repeat domains 2	2	0.8
KANSL1	KAT8 regulatory NSL complex subunit 1	2	0.8
KAT2A	K(lysine) acetyltransferase 2A	4	1.5
KAT2B	K(lysine) acetyltransferase 2B	3	1.1
KAT7	K(lysine) acetyltransferase 7	1	0.4
KATNAL1	katanin p60 subunit A-like 1	4	1.5
KCTD12	potassium channel tetramerization domain containing 12	1	0.4
KCTD2	potassium channel tetramerization domain containing 2	1	0.4
KCTD20	potassium channel tetramerization domain containing 20	1	0.4
KCTD5	potassium channel tetramerization domain containing 5	1	0.4
KCTD6	potassium channel tetramerization domain containing 6	1	0.4
KDM2A	lysine (K)-specific demethylase 2A	1	0.4
KDM3B	lysine (K)-specific demethylase 3B	1	0.4
KDM4A	lysine (K)-specific demethylase 4A	2	0.8
KDM5B	lysine (K)-specific demethylase 5B	2	0.8
KDM5D	lysine (K)-specific demethylase 5D	1	0.4
KHDRBS1	KH domain containing, RNA binding, signal transduction associated 1	1	0.4
KIAA0100	KIAA0100	2	0.8
KIF1A	kinesin family member 1A	1	0.4
KIF1B	kinesin family member 1B	2	0.8
KIF1C	kinesin family member 1C	1	0.4
KIF2C	kinesin family member 2C	2	0.8
KIF3B	kinesin family member 3B	4	1.5
KLC1	kinesin light chain 1	1	0.4
KLC4	kinesin light chain 4	4	1.5
KLHL42	kelch-like family member 42	3	1.1
KPNA2	karyopherin alpha 2 (RAG cohort 1, importin alpha 1)	3	1.1
KPNB1	karyopherin (importin) beta 1	5	1.9

KYNU	kynureninase	1	0.4
LAMC1	laminin, gamma 1 (formerly LAMB2)	1	0.4
LARP1	La ribonucleoprotein domain family, member 1	2	0.8
LARS	leucyl-tRNA synthetase	1	0.4
LATS2	large tumor suppressor kinase 2	4	1.5
LBR	lamin B receptor	2	0.8
LDHB	lactate dehydrogenase B	2	0.8
LEPRE1	leucine proline-enriched proteoglycan (leprecan) 1	1	0.4
LEPREL2	leprecan-like 2	1	0.4
LEPROTL1	leptin receptor overlapping transcript-like 1	1	0.4
LGALS3	lectin, galactoside-binding, soluble, 3	1	0.4
LGALS3BP	lectin, galactoside-binding, soluble, 3 binding protein	1	0.4
LGR4	leucine-rich repeat containing G protein-coupled receptor 4	1	0.4
LIMCH1	LIM and calponin homology domains 1	1	0.4
LIMK1	LIM domain kinase 1	6	2.3
LIN7C	lin-7 homolog C (C. elegans)	1	0.4
LMAN1	lectin, mannose-binding, 1	2	0.8
LMCD1	LIM and cysteine-rich domains 1	1	0.4
LMNB1	lamin B1	2	0.8
LONRF1	LON peptidase N-terminal domain and ring finger 1	18	6.9
LPCAT1	lysophosphatidylcholine acyltransferase 1	4	1.5
LRIG1	leucine-rich repeats and immunoglobulin-like domains 1	1	0.4
LRRC8A	leucine rich repeat containing 8 family, member A	2	0.8
LUC7L3	LUC7-like 3 (S. cerevisiae)	1	0.4
MACF1	microtubule-actin crosslinking factor 1	1	0.4
MAGI1	membrane associated guanylate kinase, WW and PDZ domain containing 1	1	0.4
MAP1B	microtubule-associated protein 1B	5	1.9
MAP2	microtubule-associated protein 2	1	0.4
MAP3K10	mitogen-activated protein kinase kinase kinase 10	7	2.7
MAP3K11	mitogen-activated protein kinase kinase kinase 11	2	0.8
MAP3K5	mitogen-activated protein kinase kinase kinase 5	1	0.4
MAPK1	mitogen-activated protein kinase 1	1	0.4
MAPK13	mitogen-activated protein kinase 13	1	0.4
MAPKAPK2	mitogen-activated protein kinase-activated protein kinase 2	3	1.1
MAPRE1	microtubule-associated protein, RP/EB family, member 1	1	0.4
MAPRE2	microtubule-associated protein, RP/EB family, member 2	2	0.8
MARK4	MAP/microtubule affinity-regulating kinase 4	1	0.4
MBNL1	muscleblind-like splicing regulator 1	1	0.4

MCL1	myeloid cell leukemia 1	1	0.4
MCM2	minichromosome maintenance complex component 2	1	0.4
MCM3	minichromosome maintenance complex component 3	5	1.9
MCM7	minichromosome maintenance complex component 7	28	10.7
MCU	mitochondrial calcium uniporter	1	0.4
MECP2	methyl CpG binding protein 2	1	0.4
MED1	mediator complex subunit 1	2	0.8
MED13	mediator complex subunit 13	6	2.3
MED24	mediator complex subunit 24	11	4.2
MEF2A	myocyte enhancer factor 2A	1	0.4
MEF2C	myocyte enhancer factor 2C	5	1.9
MEF2D	myocyte enhancer factor 2D	7	2.7
MFHAS1	malignant fibrous histiocytoma amplified sequence 1	2	0.8
MGRN1	mahogunin ring finger 1, E3 ubiquitin protein ligase	2	0.8
MIB1	mindbomb E3 ubiquitin protein ligase 1	1	0.4
MID1	midline 1	1	0.4
MMP14	matrix metallopeptidase 14 (membrane-inserted)	1	0.4
MMP2	matrix metallopeptidase 2 (gelatinase A, 72kDa gelatinase, 72kDa type IV collagenase)	1	0.4
MMP9	matrix metallopeptidase 9 (gelatinase B, 92kDa gelatinase, 92kDa type IV collagenase)	11	4.2
MPDZ	multiple PDZ domain protein	1	0.4
MPP3	membrane protein, palmitoylated 3 (MAGUK p55 subfamily member 3)	4	1.5
MPP6	membrane protein, palmitoylated 6 (MAGUK p55 subfamily member 6)	1	0.4
MTA2	metastasis associated 1 family, member 2	1	0.4
MTMR12	myotubularin related protein 12	5	1.9
MTMR3	myotubularin related protein 3	2	0.8
MTMR4	myotubularin related protein 4	1	0.4
MTMR9	myotubularin related protein 9	3	1.1
MTOR	mechanistic target of rapamycin (serine/threonine kinase)	1	0.4
MTUS1	microtubule associated tumor suppressor 1	2	0.8
MYH10	myosin, heavy chain 10, non-muscle	2	0.8
MYO10	myosin X	1	0.4
MYO1B	myosin IB	1	0.4
MYO1C	myosin IC	2	0.8
MYO1D	myosin ID	4	1.5
MYO5A	myosin VA (heavy chain 12, myoxin)	3	1.1
MYO6	myosin VI	1	0.4
MYO9A	myosin IXA	1	0.4
NAA15	N(alpha)-acetyltransferase 15, NatA auxiliary subunit	2	0.8

NABP2	nucleic acid binding protein 2	2	0.8
NARF	nuclear prelamin A recognition factor	1	0.4
NBEA	neurobeachin	17	6.5
NCBP2	nuclear cap binding protein subunit 2, 20kDa	1	0.4
NCEH1	neutral cholesterol ester hydrolase 1	3	1.1
NCKAP1	NCK-associated protein 1	1	0.4
NCOA3	nuclear receptor coactivator 3	32	12.3
NCOA5	nuclear receptor coactivator 5	1	0.4
NCOR1	nuclear receptor corepressor 1	1	0.4
NCOR2	nuclear receptor corepressor 2	8	3.1
NDE1	nudE neurodevelopment protein 1	7	2.7
NDEL1	nudE neurodevelopment protein 1-like 1	3	1.1
NEDD4L	neural precursor cell expressed, developmentally down-regulated 4-like, E3 ubiquitin protein ligase	2	0.8
NEDD9	neural precursor cell expressed, developmentally down-regulated 9	1	0.4
NEK6	NIMA-related kinase 6	1	0.4
NEK9	NIMA-related kinase 9	1	0.4
NELFCD	negative elongation factor complex member C/D	1	0.4
NET1	neuroepithelial cell transforming 1	1	0.4
NFATC2	nuclear factor of activated T-cells, cytoplasmic, calcineurin-dependent 2	9	3.4
NFIA	nuclear factor I/A	1	0.4
NFKB1	nuclear factor of kappa light polypeptide gene enhancer in B-cells 1	2	0.8
NFKBIZ	nuclear factor of kappa light polypeptide gene enhancer in B-cells inhibitor, zeta	1	0.4
NHLRC3	NHL repeat containing 3	1	0.4
NID1	nidogen 1	1	0.4
NKIRAS2	NFKB inhibitor interacting Ras-like 2	1	0.4
NKRF	NFKB repressing factor	1	0.4
NLGN4Y	neuroligin 4, Y-linked	1	0.4
NLK	nemo-like kinase	2	0.8
NMT1	N-myristoyltransferase 1	5	1.9
NOD1	nucleotide-binding oligomerization domain containing 1	1	0.4
NOS3	nitric oxide synthase 3 (endothelial cell)	1	0.4
NPAS2	neuronal PAS domain protein 2	1	0.4
NPC1	Niemann-Pick disease, type C1	1	0.4
NR1H3	nuclear receptor subfamily 1, group H, member 3	3	1.1
NR2C2	nuclear receptor subfamily 2, group C, member 2	2	0.8
NR2F6	nuclear receptor subfamily 2, group F, member 6	1	0.4
NR3C1	nuclear receptor subfamily 3, group C, member 1 (glucocorticoid receptor)	1	0.4
NR3C2	nuclear receptor subfamily 3, group C, member 2	1	0.4

NR4A2	nuclear receptor subfamily 4, group A, member 2	3	1.1
NRP1	neuropilin 1	5	1.9
NRXN3	neurexin 3	1	0.4
NSUN2	NOP2/Sun RNA methyltransferase family, member 2	2	0.8
NUFIP2	nuclear fragile X mental retardation protein interacting protein 2	7	2.7
NUP153	nucleoporin 153kDa	2	0.8
NUP205	nucleoporin 205kDa	1	0.4
OPTN	optineurin	1	0.4
OSBPL10	oxysterol binding protein-like 10	1	0.4
OSBPL1A	oxysterol binding protein-like 1A	1	0.4
OSBPL8	oxysterol binding protein-like 8	2	0.8
OTUD4	OTU deubiquitinase 4	2	0.8
OXCT1	3-oxoacid CoA transferase 1	3	1.1
PABPC1	poly(A) binding protein, cytoplasmic 1	1	0.4
PACSIN3	protein kinase C and casein kinase substrate in neurons 3	1	0.4
PAK1	p21 protein (Cdc42/Rac)-activated kinase 1	10	3.8
PAK4	p21 protein (Cdc42/Rac)-activated kinase 4	15	5.7
PAPD7	PAP associated domain containing 7	6	2.3
PARD6B	par-6 family cell polarity regulator beta	2	0.8
PBX2	pre-B-cell leukemia homeobox 2	2	0.8
PCBP1	poly(rC) binding protein 1	1	0.4
PCBP2	poly(rC) binding protein 2	1	0.4
PCDH18	protocadherin 18	1	0.4
PCGF2	polycomb group ring finger 2	3	1.1
PCGF3	polycomb group ring finger 3	3	1.1
PCLO	piccolo presynaptic cytomatrix protein	1	0.4
PDE4D	phosphodiesterase 4D, cAMP-specific	5	1.9
PDHX	pyruvate dehydrogenase complex, component X	1	0.4
PDLIM7	PDZ and LIM domain 7 (enigma)	1	0.4
PDS5A	PDS5, regulator of cohesion maintenance, homolog A (S. cerevisiae)	2	0.8
PDS5B	PDS5, regulator of cohesion maintenance, homolog B (S. cerevisiae)	1	0.4
PEX11B	peroxisomal biogenesis factor 11 beta	2	0.8
PFAS	phosphoribosylformylglycinamidine synthase	1	0.4
PFDN2	prefoldin subunit 2	1	0.4
PFKM	phosphofructokinase, muscle	1	0.4
PFKP	phosphofructokinase, platelet	7	2.7
PGRMC1	progesterone receptor membrane component 1	1	0.4
PHC2	polyhomeotic homolog 2 (Drosophila)	1	0.4
PHF1	PHD finger protein 1	2	0.8
PHGDH	phosphoglycerate dehydrogenase	2	0.8

PHKG2	phosphorylase kinase, gamma 2 (testis)	2	0.8
PHTF1	putative homeodomain transcription factor 1	1	0.4
PIAS3	protein inhibitor of activated STAT, 3	1	0.4
PIGS	phosphatidylinositol glycan anchor biosynthesis, class S	2	0.8
PIP5K1A	phosphatidylinositol-4-phosphate 5-kinase, type I, alpha	1	0.4
PITPNA	phosphatidylinositol transfer protein, alpha	2	0.8
PKD2	polycystic kidney disease 2 (autosomal dominant)	1	0.4
PKN3	protein kinase N3	1	0.4
PKP4	plakophilin 4	1	0.4
PLAT	plasminogen activator, tissue	4	1.5
PLAU	plasminogen activator, urokinase	1	0.4
PLD3	phospholipase D family, member 3	1	0.4
PLEKHA5	pleckstrin homology domain containing, family A member 5	1	0.4
PLEKHF2	pleckstrin homology domain containing, family F (with FYVE domain) member 2	1	0.4
PLEKHG2	pleckstrin homology domain containing, family G (with RhoGef domain) member 2	3	1.1
PLOD1	procollagen-lysine, 2-oxoglutarate 5-dioxygenase 1	1	0.4
PLOD2	procollagen-lysine, 2-oxoglutarate 5-dioxygenase 2	3	1.1
PLOD3	procollagen-lysine, 2-oxoglutarate 5-dioxygenase 3	18	6.9
PLS1	plastin 1	1	0.4
PNN	pinin, desmosome associated protein	3	1.1
PNPLA7	patatin-like phospholipase domain containing 7	1	0.4
POGK	pogo transposable element with KRAB domain	1	0.4
POGZ	pogo transposable element with ZNF domain	11	4.2
POLR2A	polymerase (RNA) II (DNA directed) polypeptide A, 220kDa	1	0.4
POLR2L	polymerase (RNA) II (DNA directed) polypeptide L, 7.6kDa	1	0.4
POU2F1	POU class 2 homeobox 1	1	0.4
PPARA	peroxisome proliferator-activated receptor alpha	2	0.8
PPARD	peroxisome proliferator-activated receptor delta	2	0.8
PPFIA1	protein tyrosine phosphatase, receptor type, f polypeptide (PTPRF), interacting protein (liprin), alpha 1	1	0.4
PPFIBP1	PTPRF interacting protein, binding protein 1 (liprin beta 1)	1	0.4
PPIA	peptidylprolyl isomerase A (cyclophilin A)	2	0.8
PPIB	peptidylprolyl isomerase B (cyclophilin B)	1	0.4
PPIF	peptidylprolyl isomerase F	1	0.4
PPL	periplakin	1	0.4
PPP1R12A	protein phosphatase 1, regulatory subunit 12A	3	1.1
PPP1R12C	protein phosphatase 1, regulatory subunit 12C	1	0.4
PPP2R5D	protein phosphatase 2, regulatory subunit B,	1	0.4

	delta		
PPP6R1	protein phosphatase 6, regulatory subunit 1	2	0.8
PPP6R3	protein phosphatase 6, regulatory subunit 3	12	4.6
PRKAA1	protein kinase, AMP-activated, alpha 1 catalytic subunit	3	1.1
PRKCI	protein kinase C, iota	1	0.4
PRKD3	protein kinase D3	1	0.4
PRKG1	protein kinase, cGMP-dependent, type I	1	0.4
PRPF8	pre-mRNA processing factor 8	3	1.1
PRRC2A	proline-rich coiled-coil 2A	19	7.3
PRRC2B	proline-rich coiled-coil 2B	1	0.4
PRUNE2	prune homolog 2 (Drosophila)	1	0.4
PSD3	pleckstrin and Sec7 domain containing 3	1	0.4
PSMB5	proteasome (prosome, macropain) subunit, beta type, 5	4	1.5
PSMD11	proteasome (prosome, macropain) 26S subunit, non-ATPase, 11	3	1.1
PSMD13	proteasome (prosome, macropain) 26S subunit, non-ATPase, 13	1	0.4
PSMD3	proteasome (prosome, macropain) 26S subunit, non-ATPase, 3	3	1.1
PSMD6	proteasome (prosome, macropain) 26S subunit, non-ATPase, 6	2	0.8
PSME3	proteasome (prosome, macropain) activator subunit 3 (PA28 gamma; Ki)	1	0.4
PSPC1	paraspeckle component 1	1	0.4
PSTPIP1	proline-serine-threonine phosphatase interacting protein 1	1	0.4
PTBP1	polypyrimidine tract binding protein 1	2	0.8
PTBP3	polypyrimidine tract binding protein 3	1	0.4
PTGS2	prostaglandin-endoperoxide synthase 2 (prostaglandin G/H synthase and cyclooxygenase)	1	0.4
PTK2	protein tyrosine kinase 2	14	5.4
PTK2B	protein tyrosine kinase 2 beta	1	0.4
PTP4A3	protein tyrosine phosphatase type IVA, member 3	1	0.4
PTPN1	protein tyrosine phosphatase, non-receptor type 1	18	6.9
PTPN12	protein tyrosine phosphatase, non-receptor type 12	9	3.4
PTPN3	protein tyrosine phosphatase, non-receptor type 3	1	0.4
PTPRD	protein tyrosine phosphatase, receptor type, D	1	0.4
PTPRF	protein tyrosine phosphatase, receptor type, F	1	0.4
PTPRJ	protein tyrosine phosphatase, receptor type, J	6	2.3
PUM1	pumilio RNA-binding family member 1	3	1.1
QKI	QKI, KH domain containing, RNA binding	1	0.4
QRICH2	glutamine rich 2	1	0.4
QSOX2	quiescin Q6 sulfhydryl oxidase 2	1	0.4
RAB22A	RAB22A, member RAS oncogene family	8	3.1

RAB34	RAB34, member RAS oncogene family	4	1.5
RABGAP1	RAB GTPase activating protein 1	1	0.4
RABGAP1L	RAB GTPase activating protein 1-like	1	0.4
RABL6	RAB, member RAS oncogene family-like 6	1	0.4
RAD23B	RAD23 homolog B (S. cerevisiae)	1	0.4
RAD9A	RAD9 homolog A (S. pombe)	2	0.8
RAE1	ribonucleic acid export 1	1	0.4
RAI14	retinoic acid induced 14	2	0.8
RALY	RALY heterogeneous nuclear ribonucleoprotein	3	1.1
RANBP10	RAN binding protein 10	1	0.4
RANBP9	RAN binding protein 9	1	0.4
RAP1B	RAP1B, member of RAS oncogene family	1	0.4
RAPGEF2	Rap guanine nucleotide exchange factor (GEF) 2	1	0.4
RAPGEF4	Rap guanine nucleotide exchange factor (GEF) 4	1	0.4
RARS	arginyl-tRNA synthetase	2	0.8
RASAL2	RAS protein activator like 2	1	0.4
RASSF8	Ras association (RalGDS/AF-6) domain family (N-terminal) member 8	1	0.4
RBBP8	retinoblastoma binding protein 8	4	1.5
RBL1	retinoblastoma-like 1	11	4.2
RBL2	retinoblastoma-like 2	2	0.8
RBM10	RNA binding motif protein 10	1	0.4
RBM12B	RNA binding motif protein 12B	3	1.1
RBM5	RNA binding motif protein 5	1	0.4
RBM6	RNA binding motif protein 6	1	0.4
RBM8A	RNA binding motif protein 8A	1	0.4
RELA	v-rel avian reticuloendotheliosis viral oncogene homolog A	1	0.4
RELB	v-rel avian reticuloendotheliosis viral oncogene homolog B	2	0.8
REPS1	RALBP1 associated Eps domain containing 1	7	2.7
RERE	arginine-glutamic acid dipeptide (RE) repeats	2	0.8
RFC3	replication factor C (activator 1) 3, 38kDa	1	0.4
RFFL	ring finger and FYVE-like domain containing E3 ubiquitin protein ligase	1	0.4
RGS19	regulator of G-protein signaling 19	3	1.1
RHOT1	ras homolog family member T1	2	0.8
RICTOR	RPTOR independent companion of MTOR, complex 2	1	0.4
RIMS1	regulating synaptic membrane exocytosis 1	1	0.4
RIMS2	regulating synaptic membrane exocytosis 2	1	0.4
RIN3	Ras and Rab interactor 3	1	0.4
RIPK2	receptor-interacting serine-threonine kinase 2	1	0.4
RIPK3	receptor-interacting serine-threonine kinase 3	1	0.4
RLF	rearranged L-myc fusion	2	0.8
RNF111	ring finger protein 111	2	0.8

RNF114	ring finger protein 114	1	0.4
RNF149	ring finger protein 149	1	0.4
RNF167	ring finger protein 167	1	0.4
RNF40	ring finger protein 40, E3 ubiquitin protein ligase	4	1.5
RNF5	ring finger protein 5, E3 ubiquitin protein ligase	1	0.4
RNH1	ribonuclease/angiogenin inhibitor 1	1	0.4
ROCK1	Rho-associated, coiled-coil containing protein kinase 1	15	5.7
RPIA	ribose 5-phosphate isomerase A	1	0.4
RPL23	ribosomal protein L23	1	0.4
RPL28	ribosomal protein L28	1	0.4
RPRD1A	regulation of nuclear pre-mRNA domain containing 1A	2	0.8
RPRD1B	regulation of nuclear pre-mRNA domain containing 1B	5	1.9
RPS2	ribosomal protein S2	1	0.4
RPS24	ribosomal protein S24	2	0.8
RPS27	ribosomal protein S27	2	0.8
RPS6KA5	ribosomal protein S6 kinase, 90kDa, polypeptide 5	1	0.4
RRBP1	ribosome binding protein 1	1	0.4
RTN4	reticulon 4	2	0.8
RUNDC3B	RUN domain containing 3B	1	0.4
RUNX2	runt-related transcription factor 2	7	2.7
RXRB	retinoid X receptor, beta	10	3.8
SCAMP1	secretory carrier membrane protein 1	1	0.4
SCARB1	scavenger receptor class B, member 1	2	0.8
SCD	stearoyl-CoA desaturase (delta-9-desaturase)	1	0.4
SCRIB	scribbled planar cell polarity protein	19	7.3
SCRN1	secernin 1	1	0.4
SEC23A	Sec23 homolog A (S. cerevisiae)	1	0.4
SEC23IP	SEC23 interacting protein	1	0.4
SEC24B	SEC24 family member B	1	0.4
SEC24C	SEC24 family member C	4	1.5
SEC31A	SEC31 homolog A (S. cerevisiae)	1	0.4
SENP1	SUMO1/sentrin specific peptidase 1	4	1.5
SENP2	SUMO1/sentrin/SMT3 specific peptidase 2	2	0.8
SENP3	SUMO1/sentrin/SMT3 specific peptidase 3	2	0.8
SEPT2	septin 2	1	0.4
SERINC5	serine incorporator 5	1	0.4
SERPINE1	serpin peptidase inhibitor, clade E (nexin, plasminogen activator inhibitor type 1), member 1	4	1.5
SETD5	SET domain containing 5	4	1.5
SETD7	SET domain containing (lysine methyltransferase) 7	1	0.4

SF3A1	splicing factor 3a, subunit 1, 120kDa	1	0.4
SF3B2	splicing factor 3b, subunit 2, 145kDa	1	0.4
SF3B3	splicing factor 3b, subunit 3, 130kDa	1	0.4
SFXN1	sideroflexin 1	1	0.4
SGK1	serum/glucocorticoid regulated kinase 1	1	0.4
SH2D4A	SH2 domain containing 4A	1	0.4
SH3D19	SH3 domain containing 19	1	0.4
SH3PXD2A	SH3 and PX domains 2A	1	0.4
SIK2	salt-inducible kinase 2	1	0.4
SIPA1L3	signal-induced proliferation-associated 1 like 3	1	0.4
SIRT1	sirtuin 1	1	0.4
SKI	SKI proto-oncogene	3	1.1
SLC13A3	solute carrier family 13 (sodium-dependent dicarboxylate transporter), member 3	2	0.8
SLC16A1	solute carrier family 16 (monocarboxylate transporter), member 1	1	0.4
SLC19A2	solute carrier family 19 (thiamine transporter), member 2	1	0.4
SLC25A13	solute carrier family 25 (aspartate/glutamate carrier), member 13	10	3.8
SLC25A40	solute carrier family 25, member 40	4	1.5
SLC25A6	solute carrier family 25 (mitochondrial carrier; adenine nucleotide translocator), member 6	1	0.4
SLC2A4	solute carrier family 2 (facilitated glucose transporter), member 4	3	1.1
SLC39A1	solute carrier family 39 (zinc transporter), member 1	7	2.7
SLC39A10	solute carrier family 39 (zinc transporter), member 10	3	1.1
SLC39A14	solute carrier family 39 (zinc transporter), member 14	1	0.4
SLC39A8	solute carrier family 39 (zinc transporter), member 8	1	0.4
SLC46A3	solute carrier family 46, member 3	3	1.1
SLC7A1	solute carrier family 7 (cationic amino acid transporter, y+ system), member 1	7	2.7
SLC7A2	solute carrier family 7 (cationic amino acid transporter, y+ system), member 2	1	0.4
SLC9A3R1	solute carrier family 9, subfamily A (NHE3, cation proton antiporter 3), member 3 regulator 1	1	0.4
SLC9A6	solute carrier family 9, subfamily A (NHE6, cation proton antiporter 6), member 6	2	0.8
SMARCA2	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily a, member 2	2	0.8
SMARCC2	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily c, member 2	5	1.9
SMC1A	structural maintenance of chromosomes 1A	1	0.4
SMCHD1	structural maintenance of chromosomes flexible hinge domain containing 1	2	0.8
SMG1	SMG1 phosphatidylinositol 3-kinase-related kinase	1	0.4
SMOX	spermine oxidase	1	0.4

SMYD2	SET and MYND domain containing 2	1	0.4
SMYD3	SET and MYND domain containing 3	2	0.8
SNAP25	synaptosomal-associated protein, 25kDa	1	0.4
SNAP29	synaptosomal-associated protein, 29kDa	1	0.4
SNRPA	small nuclear ribonucleoprotein polypeptide A	1	0.4
SNX1	sorting nexin 1	1	0.4
SORBS3	sorbin and SH3 domain containing 3	1	0.4
SOS2	son of sevenless homolog 2 (Drosophila)	1	0.4
SOX5	SRY (sex determining region Y)-box 5	14	5.4
SP1	Sp1 transcription factor	1	0.4
SP4	Sp4 transcription factor	1	0.4
SPAG9	sperm associated antigen 9	5	1.9
SPEN	spen family transcriptional repressor	1	0.4
SPNS1	spinster homolog 1 (Drosophila)	1	0.4
SPRY1	sprouty homolog 1, antagonist of FGF signaling (Drosophila)	1	0.4
SPTBN1	spectrin, beta, non-erythrocytic 1	7	2.7
SRCAP	Snf2-related CREBBP activator protein	1	0.4
SRF	serum response factor (c-fos serum response element-binding transcription factor)	5	1.9
SRGAP1	SLIT-ROBO Rho GTPase activating protein 1	2	0.8
SRGAP2	SLIT-ROBO Rho GTPase activating protein 2	2	0.8
SRI	sorcin	8	3.1
SRSF1	serine/arginine-rich splicing factor 1	1	0.4
SRSF4	serine/arginine-rich splicing factor 4	1	0.4
SSRP1	structure specific recognition protein 1	1	0.4
STAG1	stromal antigen 1	1	0.4
STAM	signal transducing adaptor molecule (SH3 domain and ITAM motif) 1	2	0.8
STAMBPL1	STAM binding protein-like 1	1	0.4
STAT1	signal transducer and activator of transcription 1, 91kDa	1	0.4
STK24	serine/threonine kinase 24	1	0.4
STK3	serine/threonine kinase 3	14	5.4
STK35	serine/threonine kinase 35	1	0.4
STK4	serine/threonine kinase 4	10	3.8
STOM	stomatin	1	0.4
STRN	striatin, calmodulin binding protein	3	1.1
STX1A	syntaxin 1A (brain)	6	2.3
STX7	syntaxin 7	1	0.4
STXBP1	syntaxin binding protein 1	4	1.5
STXBP3	syntaxin binding protein 3	1	0.4
SUPT5H	suppressor of Ty 5 homolog (S. cerevisiae)	1	0.4
SYNE1	spectrin repeat containing, nuclear envelope 1	1	0.4
SYNJ1	synaptojanin 1	1	0.4

SYPL1	synaptophysin-like 1	1	0.4
TADA2A	transcriptional adaptor 2A	1	0.4
TAF4	TAF4 RNA polymerase II, TATA box binding protein (TBP)-associated factor, 135kDa	1	0.4
TARS	threonyl-tRNA synthetase	3	1.1
TAX1BP1	Tax1 (human T-cell leukemia virus type I) binding protein 1	1	0.4
TBC1D22B	TBC1 domain family, member 22B	2	0.8
TBC1D4	TBC1 domain family, member 4	5	1.9
TBC1D9	TBC1 domain family, member 9 (with GRAM domain)	1	0.4
TCEA2	transcription elongation factor A (SII), 2	4	1.5
TEK	TEK tyrosine kinase, endothelial	1	0.4
TES	testis derived transcript (3 LIM domains)	7	2.7
TESK1	testis-specific kinase 1	3	1.1
TGFB1I1	transforming growth factor beta 1 induced transcript 1	1	0.4
TGFBR2	transforming growth factor, beta receptor II (70/80kDa)	1	0.4
TGM2	transglutaminase 2	2	0.8
THAP1	THAP domain containing, apoptosis associated protein 1	1	0.4
THBS1	thrombospondin 1	3	1.1
THOP1	thimet oligopeptidase 1	3	1.1
THRA	thyroid hormone receptor, alpha	20	7.7
TIAM1	T-cell lymphoma invasion and metastasis 1	6	2.3
TIMM50	translocase of inner mitochondrial membrane 50 homolog (S. cerevisiae)	2	0.8
TJP2	tight junction protein 2	1	0.4
TLE4	transducin-like enhancer of split 3	2	0.8
TLE4	transducin-like enhancer of split 4	1	0.4
TLN1	talin 1	3	1.1
TLR2	toll-like receptor 2	1	0.4
TLR4	toll-like receptor 4	5	1.9
TMEM14A	transmembrane protein 14A	2	0.8
TMEM170B	transmembrane protein 170B	1	0.4
TMEM2	transmembrane protein 2	1	0.4
TMEM30A	transmembrane protein 30A	2	0.8
TMEM63B	transmembrane protein 63B	1	0.4
TMEM87A	transmembrane protein 87A	1	0.4
TNFAIP1	tumor necrosis factor, alpha-induced protein 1 (endothelial)	9	3.4
TNKS	tankyrase, TRF1-interacting ankyrin-related ADP-ribose polymerase	3	1.1
TNPO1	transportin 1	1	0.4
TNRC6B	trinucleotide repeat containing 6B	3	1.1
TNS3	tensin 3	1	0.4
TOB1	transducer of ERBB2, 1	7	2.7

TOB2	transducer of ERBB2, 2	1	0.4
TOM1	target of myb1 (chicken)	1	0.4
TOMM34	translocase of outer mitochondrial membrane 34	33	12.6
TPD52	tumor protein D52	1	0.4
TPD52L2	tumor protein D52-like 2	1	0.4
TRA2B	transformer 2 beta homolog (Drosophila)	4	1.5
TRAF2	TNF receptor-associated factor 2	1	0.4
TRAF4	TNF receptor-associated factor 4	1	0.4
TRAP1	TNF receptor-associated protein 1	1	0.4
TRAPPC3	trafficking protein particle complex 3	2	0.8
TRERF1	transcriptional regulating factor 1	1	0.4
TRIB1	tribbles pseudokinase 1	4	1.5
TRIB3	tribbles pseudokinase 3	6	2.3
TRIM2	tripartite motif containing 2	3	1.1
TRIM21	tripartite motif containing 21	1	0.4
TRIM28	tripartite motif containing 28	10	3.8
TRIM37	tripartite motif containing 37	1	0.4
TRIM39	tripartite motif containing 39	5	1.9
TRIM69	tripartite motif containing 69	1	0.4
TRIP12	thyroid hormone receptor interactor 12	3	1.1
TRIP6	thyroid hormone receptor interactor 6	4	1.5
TRMT2A	tRNA methyltransferase 2 homolog A (S. cerevisiae)	1	0.4
TSC22D4	TSC22 domain family, member 4	2	0.8
TTC37	tetratricopeptide repeat domain 37	1	0.4
TUBA1C	tubulin, alpha 1c	1	0.4
TUBB	tubulin, beta class I	6	2.3
TUBB4B	tubulin, beta 4B class IVb	3	1.1
TULP3	tubby like protein 3	1	0.4
TXLNA	taxilin alpha	4	1.5
TXNIP	thioredoxin interacting protein	2	0.8
TYMS	thymidylate synthetase	3	1.1
U2AF2	U2 small nuclear RNA auxiliary factor 2	1	0.4
UBAP2	ubiquitin associated protein 2	1	0.4
UBAP2L	ubiquitin associated protein 2-like	5	1.9
UBASH3B	ubiquitin associated and SH3 domain containing B	1	0.4
UBE2R2	ubiquitin-conjugating enzyme E2R 2	3	1.1
UBR1	ubiquitin protein ligase E3 component n-recognin 1	2	0.8
UCHL3	ubiquitin carboxyl-terminal esterase L3 (ubiquitin thiolesterase)	4	1.5
UFM1	ubiquitin-fold modifier 1	1	0.4
UGGT1	UDP-glucose glycoprotein glucosyltransferase 1	1	0.4
ULK1	unc-51 like autophagy activating kinase 1	2	0.8

UNC119	unc-119 homolog (C. elegans)	2	0.8
UNC93B1	unc-93 homolog B1 (C. elegans)	5	1.9
USF1	upstream transcription factor 1	1	0.4
USF2	upstream transcription factor 2, c-fos interacting	4	1.5
USP10	ubiquitin specific peptidase 10	1	0.4
USP15	ubiquitin specific peptidase 15	1	0.4
USP2	ubiquitin specific peptidase 2	2	0.8
USP21	ubiquitin specific peptidase 21	6	2.3
USP22	ubiquitin specific peptidase 22	2	0.8
USP25	ubiquitin specific peptidase 25	6	2.3
USP28	ubiquitin specific peptidase 28	1	0.4
USP31	ubiquitin specific peptidase 31	5	1.9
USP36	ubiquitin specific peptidase 36	3	1.1
USP45	ubiquitin specific peptidase 45	1	0.4
USP49	ubiquitin specific peptidase 49	1	0.4
USP5	ubiquitin specific peptidase 5 (isopeptidase T)	7	2.7
USP7	ubiquitin specific peptidase 7 (herpes virus-associated)	1	0.4
USP9Y	ubiquitin specific peptidase 9, Y-linked	1	0.4
UTP20	UTP20, small subunit (SSU) processome component, homolog (yeast)	1	0.4
UTRN	utrophin	2	0.8
UTY	ubiquitously transcribed tetratricopeptide repeat containing, Y-linked	1	0.4
VAC14	Vac14 homolog (S. cerevisiae)	1	0.4
VAMP3	vesicle-associated membrane protein 3	2	0.8
VAPB	VAMP (vesicle-associated membrane protein)-associated protein B and C	26	10
VARS	valyl-tRNA synthetase	3	1.1
VASP	vasodilator-stimulated phosphoprotein	4	1.5
VAV2	vav 2 guanine nucleotide exchange factor	1	0.4
VCP	valosin containing protein	1	0.4
VDAC2	voltage-dependent anion channel 2	2	0.8
VDAC3	voltage-dependent anion channel 3	1	0.4
VEGFA	vascular endothelial growth factor A	23	8.8
VIM	vimentin	2	0.8
VPS13C	vacuolar protein sorting 13 homolog C (S. cerevisiae)	3	1.1
VPS37B	vacuolar protein sorting 37 homolog B (S. cerevisiae)	2	0.8
WBP2	WW domain binding protein 2	1	0.4
WDR17	WD repeat domain 17	1	0.4
WDR62	WD repeat domain 62	6	2.3
WDR78	WD repeat domain 78	1	0.4
WEE1	WEE1 G2 checkpoint kinase	2	0.8
WIPF2	WAS/WASL interacting protein family, member 2	1	0.4

WNK1	WNK lysine deficient protein kinase 1	1	0.4
XPO7	exportin 7	2	0.8
XYLT2	xylosyltransferase II	2	0.8
YES1	YES proto-oncogene 1, Src family tyrosine kinase	3	1.1
YY1	YY1 transcription factor	1	0.4
ZBED1	zinc finger, BED-type containing 1	2	0.8
ZBTB17	zinc finger and BTB domain containing 17	1	0.4
ZBTB18	zinc finger and BTB domain containing 18	1	0.4
ZBTB34	zinc finger and BTB domain containing 34	1	0.4
ZBTB7B	zinc finger and BTB domain containing 7B	6	2.3
ZBTB9	zinc finger and BTB domain containing 9	3	1.1
ZC3H12C	zinc finger CCCH-type containing 12C	2	0.8
ZDHHC7	zinc finger, DHHC-type containing 7	1	0.4
ZEB2	zinc finger E-box binding homeobox 2	1	0.4
ZFR	zinc finger RNA binding protein	1	0.4
ZFY	zinc finger protein, Y-linked	1	0.4
ZFYVE9	zinc finger, FYVE domain containing 9	1	0.4
ZMIZ1	zinc finger, MIZ-type containing 1	1	0.4
ZNF16	zinc finger protein 16	3	1.1
ZNF250	zinc finger protein 250	1	0.4
ZNF274	zinc finger protein 274	1	0.4
ZNF408	zinc finger protein 408	1	0.4
ZNF687	zinc finger protein 687	1	0.4
ZYX	zyxin	2	0.8

7.3 Gene set enrichment analysis of OAC drivers

Shown are the 212 pathways enriched in drivers (FDR <0.01). For each pathway, the number of genes and samples, the p-value of one-tailed hypergeometric test and the False Discovey Rate (FDR) using the Benjamini and Hochberg method are reported. Pathway size refers to the total number of genes in the pathway. Universal pathways are those with at least one perturbed driver in at least 50% of samples.

Pathway	Pathway size	Universal	Genes (n)	Samples (n)	p-value	FDR
Downstream signal transduction	348	Y	35	240	3.72E-24	4.27E-21
NGF signalling via TRKA from the plasma membrane	382	Y	36	240	7.39E-24	4.27E-21
DAP12 signaling	351	Y	34	240	6.07E-23	1.75E-20
Fc epsilon receptor (FCERI) signaling	405	Y	36	241	5.49E-23	1.75E-20
DAP12 interactions	366	Y	34	240	2.37E-22	5.47E-20
Signaling by the B Cell Receptor (BCR)	242	Y	29	240	2.87E-22	5.52E-20
IGF1R signaling cascade	292	Y	30	183	4.72E-21	5.09E-19
Insulin receptor signalling cascade	291	Y	30	183	4.28E-21	5.09E-19
IRS-mediated signalling	288	Y	30	183	3.16E-21	5.09E-19
IRS-related events triggered by IGF1R	292	Y	30	183	4.72E-21	5.09E-19
Signaling by Interleukins	402	Y	34	194	4.85E-21	5.09E-19
Downstream signaling events of B Cell Receptor (BCR)	198	Y	25	239	7.10E-20	6.84E-18
PI3K/AKT activation	130	Y	21	237	3.97E-19	3.53E-17
Interleukin-3, 5 and GM-CSF signaling	265	Y	27	182	6.25E-19	5.15E-17
PI3K events in ERBB4 signaling	127	Y	20	236	4.84E-18	3.73E-16
GAB1 signalosome	130	Y	20	236	7.81E-18	5.64E-16
VEGFA-VEGFR2 Pathway	324	Y	28	187	1.08E-17	7.32E-16
Interleukin-2 signaling	256	Y	25	180	3.74E-17	2.40E-15
PI3K/AKT Signaling in Cancer	90	Y	17	147	6.71E-17	4.08E-15
RET signaling	264	Y	25	180	7.80E-17	4.51E-15
Interleukin receptor SHC signaling	249	Y	24	180	2.24E-16	1.23E-14
ARMS-mediated activation	243	Y	23	175	1.46E-15	7.33E-14
Signalling to p38 via RIT and RIN	243	Y	23	175	1.46E-15	7.33E-14
Frs2-mediated activation	244	Y	23	175	1.60E-15	7.69E-14
MAPK1/MAPK3	245	Y	23	172	1.75E-15	7.76E-14

signaling						
Role of LAT2/NTAL/LAB on calcium mobilization	170	Y	20	236	1.69E-15	7.76E-14
Prolonged ERK activation events	246	Y	23	175	1.91E-15	8.17E-14
Signalling to RAS	250	Y	23	175	2.71E-15	1.12E-13
Negative regulation of the PI3K/AKT network	94	Y	16	135	3.21E-15	1.28E-13
Signalling to ERKs	257	Y	23	175	4.95E-15	1.90E-13
FCERI mediated MAPK activation	296	Y	24	175	1.12E-14	3.59E-13
GRB2 events in EGFR signaling	239	Y	22	172	1.12E-14	3.59E-13
RAF/MAP kinase cascade	239	Y	22	172	1.12E-14	3.59E-13
SHC1 events in EGFR signaling	239	Y	22	172	1.12E-14	3.59E-13
SHC1 events in ERBB4 signaling	239	Y	22	172	1.12E-14	3.59E-13
SOS-mediated signalling	239	Y	22	172	1.12E-14	3.59E-13
NCAM signaling for neurite out-growth	270	Y	23	173	1.44E-14	4.49E-13
VEGFR2 mediated cell proliferation	252	Y	22	172	3.35E-14	1.02E-12
PI5P, PP2A and IER3 Regulate PI3K/AKT Signaling	87	N	14	126	4.06E-13	1.20E-11
TCF dependent signaling in response to WNT	202	Y	19	143	5.35E-13	1.55E-11
Transcriptional Regulation by TP53	366	Y	24	238	1.16E-12	3.27E-11
Signaling by FGFR in disease	63	N	12	92	2.47E-12	6.80E-11
Constitutive Signaling by Aberrant PI3K in Cancer	65	N	12	123	3.66E-12	9.84E-11
Signaling by FGFR1 in disease	38	N	10	84	5.35E-12	1.40E-10
Signaling by FGFR3	40	N	10	79	9.41E-12	2.42E-10
Formation of the beta-catenin:TCF transactivating complex	60	N	11	112	3.22E-11	8.09E-10
Gastrin-CREB signalling pathway via PKC and MAPK	438	Y	24	180	5.08E-11	1.25E-09
Constitutive Signaling by EGFRvIII	15	N	7	92	8.24E-11	1.90E-09
Regulation of TP53 Degradation	35	Y	9	226	8.02E-11	1.90E-09
Signaling by EGFRvIII in Cancer	15	N	7	92	8.24E-11	1.90E-09
Regulation of TP53	36	Y	9	226	1.06E-10	2.35E-09

Expression and Degradation						
Signaling by FGFR1	50	N	10	85	1.04E-10	2.35E-09
RAF activation	25	N	8	58	1.31E-10	2.85E-09
Cyclin D associated events in G1	38	Y	9	140	1.80E-10	3.78E-09
G1 Phase	38	Y	9	140	1.80E-10	3.78E-09
Signaling by FGFR3 fusions in cancer	10	N	6	65	2.71E-10	5.58E-09
Signaling by FGFR4	41	N	9	76	3.77E-10	7.63E-09
Deactivation of the beta-catenin transactivating complex	42	N	9	54	4.75E-10	9.46E-09
Constitutive Signaling by Ligand-Responsive EGFR Cancer Variants	19	N	7	92	6.23E-10	1.18E-08
Signaling by EGFR in Cancer	19	N	7	92	6.23E-10	1.18E-08
Signaling by Ligand-Responsive EGFR Variants in Cancer	19	N	7	92	6.23E-10	1.18E-08
PI3K Cascade	79	N	11	61	7.13E-10	1.33E-08
PKMTs methylate histone lysines	47	N	9	66	1.39E-09	2.50E-08
RMTs methylate histone arginines	47	N	9	105	1.39E-09	2.50E-08
Mitotic G1-G1/S phases	141	Y	13	177	3.58E-09	6.37E-08
GPVI-mediated activation cascade	54	N	9	45	5.07E-09	8.88E-08
Oxidative Stress Induced Senescence	95	Y	11	228	5.31E-09	9.15E-08
Downstream signaling of activated FGFR3	25	N	7	66	5.63E-09	9.56E-08
Signaling by FGFR2	75	N	10	80	6.67E-09	1.12E-07
Interleukin-7 signaling	17	N	6	44	1.50E-08	2.47E-07
Signaling by cytosolic FGFR1 fusion mutants	18	N	6	45	2.23E-08	3.57E-07
Tie2 Signaling	18	N	6	63	2.23E-08	3.57E-07
Downstream signaling of activated FGFR2	30	N	7	68	2.28E-08	3.60E-07
Downstream signaling of activated FGFR1	31	N	7	72	2.92E-08	4.49E-07
FGFR1 mutant receptor activation	31	N	7	53	2.92E-08	4.49E-07
Regulation of TP53 Activity through Methylation	19	Y	6	214	3.23E-08	4.90E-07
Signaling by RAS mutants	48	N	8	55	3.70E-08	5.55E-07
Deubiquitination	280	Y	16	227	5.49E-08	8.12E-07

Interleukin-6 signaling	11	N	5	25	5.65E-08	8.16E-07
Signaling by FGFR4 in disease	11	N	5	62	5.65E-08	8.16E-07
Oncogene Induced Senescence	34	Y	7	226	5.81E-08	8.28E-07
CD209 (DC-SIGN) signaling	21	N	6	56	6.34E-08	8.93E-07
Oncogenic MAPK signaling	72	N	9	89	6.89E-08	9.58E-07
Transcriptional regulation by the AP-2 (TFAP2) family of transcription factors	35	N	7	119	7.19E-08	9.89E-07
Signaling by FGFR3 in disease	22	N	6	65	8.64E-08	1.16E-06
Signaling by FGFR3 point mutants in cancer	22	N	6	65	8.64E-08	1.16E-06
Paradoxical activation of RAF signaling by kinase inactive BRAF	38	N	7	52	1.31E-07	1.72E-06
Signaling by moderate kinase activity BRAF mutants	38	N	7	52	1.31E-07	1.72E-06
Regulation of TP53 Activity	159	Y	12	229	1.36E-07	1.76E-06
Negative regulation of MAPK pathway	40	N	7	58	1.91E-07	2.45E-06
Signaling by BRAF and RAF fusions	59	N	8	86	1.97E-07	2.50E-06
Downstream signaling of activated FGFR4	27	N	6	63	3.28E-07	4.12E-06
GRB2 events in ERBB2 signaling	16	N	5	115	5.11E-07	6.35E-06
Activation of anterior HOX genes in hindbrain development during early embryogenesis	91	N	9	64	5.33E-07	6.55E-06
NOTCH1 Intracellular Domain Regulates Transcription	47	N	7	82	6.04E-07	7.34E-06
G beta:gamma signalling through PI3Kgamma	48	N	7	44	7.01E-07	8.43E-06
G-protein beta:gamma signalling	51	N	7	44	1.07E-06	1.28E-05
Signaling by high-kinase activity BRAF mutants	34	N	6	52	1.40E-06	1.63E-05
Signaling by WNT in cancer	34	N	6	46	1.40E-06	1.63E-05
FRS-mediated FGFR3 signaling	20	N	5	50	1.75E-06	2.03E-05
Constitutive Signaling by	57	N	7	82	2.32E-06	2.55E-05

NOTCH1 HD+PEST Domain Mutants						
Constitutive Signaling by NOTCH1 PEST Domain Mutants	57	N	7	82	2.32E-06	2.55E-05
Signaling by NOTCH1 HD+PEST Domain Mutants in Cancer	57	N	7	82	2.32E-06	2.55E-05
Signaling by NOTCH1 in Cancer	57	N	7	82	2.32E-06	2.55E-05
Signaling by NOTCH1 PEST Domain Mutants in Cancer	57	N	7	82	2.32E-06	2.55E-05
MAPK3 (ERK1) activation	10	N	4	9	2.47E-06	2.69E-05
MAP2K and MAPK activation	38	N	6	52	2.78E-06	3.00E-05
SHC1 events in ERBB2 signaling	22	N	5	115	2.93E-06	3.13E-05
FRS-mediated FGFR1 signaling	23	N	5	58	3.71E-06	3.93E-05
MET activates RAS signaling	11	N	4	58	3.85E-06	4.01E-05
RHO GTPases activate IQGAPs	11	N	4	15	3.85E-06	4.01E-05
Pre-NOTCH Transcription and Translation	41	Y	6	213	4.40E-06	4.54E-05
Constitutive Signaling by AKT1 E17K in Cancer	25	N	5	50	5.76E-06	5.83E-05
FRS-mediated FGFR2 signaling	25	N	5	54	5.76E-06	5.83E-05
Signaling by FGFR2 in disease	43	N	6	67	5.86E-06	5.89E-05
PTK6 Regulates RHO GTPases, RAS GTPase and MAP kinases	26	N	5	51	7.06E-06	6.97E-05
SHC-mediated cascade:FGFR2	26	N	5	54	7.06E-06	6.97E-05
AKT phosphorylates targets in the cytosol	13	N	4	29	8.20E-06	8.03E-05
Interleukin-6 family signaling	27	N	5	25	8.60E-06	8.34E-05
Cyclin E associated events during G1/S transition	71	N	7	110	1.03E-05	9.91E-05
Signaling by NOTCH1	72	N	7	82	1.13E-05	0.000107919
Negative regulation of FGFR3 signaling	29	N	5	29	1.24E-05	0.000117663
Ub-specific processing proteases	205	Y	11	225	1.26E-05	0.0001186
Costimulation by the CD28 family	74	N	7	39	1.36E-05	0.000126363
Misspliced GSK3beta mutants stabilize beta-catenin	15	N	4	34	1.54E-05	0.000135771

phosphorylation site mutants of CTNNB1 are not targeted to the proteasome by the destruction complex	15	N	4	34	1.54E-05	0.000135771
Regulation of TP53 Activity through Acetylation	30	Y	5	203	1.48E-05	0.000135771
S33 mutants of beta-catenin aren't phosphorylated	15	N	4	34	1.54E-05	0.000135771
S37 mutants of beta-catenin aren't phosphorylated	15	N	4	34	1.54E-05	0.000135771
S45 mutants of beta-catenin aren't phosphorylated	15	N	4	34	1.54E-05	0.000135771
T41 mutants of beta-catenin aren't phosphorylated	15	N	4	34	1.54E-05	0.000135771
PI3K events in ERBB2 signaling	16	N	4	96	2.04E-05	0.000176825
Spry regulation of FGF signaling	16	N	4	26	2.04E-05	0.000176825
CD28 co-stimulation	33	N	5	38	2.40E-05	0.000205301
Negative regulation of FGFR1 signaling	33	N	5	38	2.40E-05	0.000205301
Beta-catenin phosphorylation cascade	17	N	4	34	2.64E-05	0.000222617
SUMOylation of transcription factors	17	Y	4	223	2.64E-05	0.000222617
Negative regulation of FGFR2 signaling	34	N	5	33	2.79E-05	0.000233444
Pre-NOTCH Expression and Processing	57	Y	6	213	3.09E-05	0.00025638
PI-3K cascade:FGFR3	18	N	4	31	3.37E-05	0.000275795
SHC-mediated cascade:FGFR3	18	N	4	49	3.37E-05	0.000275795
Recruitment and ATM-mediated phosphorylation of repair and signaling proteins at DNA double strand breaks	59	Y	6	205	3.76E-05	0.000306168
G1/S Transition	118	N	8	111	3.81E-05	0.000307416
DNA Double Strand Break Response	60	Y	6	205	4.15E-05	0.00033252
Toll Like Receptor 4 (TLR4) Cascade	124	N	8	28	5.43E-05	0.000432191
mTOR signalling	39	N	5	18	5.53E-05	0.000437313
C-type lectin receptors (CLRs)	125	N	8	61	5.74E-05	0.000451365
p53-Dependent G1 DNA Damage Response	64	Y	6	221	5.99E-05	0.000464611

p53-Dependent G1/S DNA damage checkpoint	64	Y	6	221	5.99E-05	0.000464611
PKB-mediated events	40	N	5	18	6.26E-05	0.000482287
PI-3K cascade:FGFR1	21	N	4	38	6.42E-05	0.000484848
SHC-mediated cascade:FGFR1	21	N	4	57	6.42E-05	0.000484848
Signal transduction by L1	21	N	4	44	6.42E-05	0.000484848
G1/S DNA Damage Checkpoints	66	Y	6	221	7.14E-05	0.000535261
CD28 dependent PI3K/Akt signaling	22	N	4	37	7.78E-05	0.00056913
CTLA4 inhibitory signaling	22	N	4	16	7.78E-05	0.00056913
FRS-mediated FGFR4 signaling	22	N	4	47	7.78E-05	0.00056913
MyD88-independent TLR3/TLR4 cascade	97	N	7	27	7.88E-05	0.00056913
Toll Like Receptor 3 (TLR3) Cascade	97	N	7	27	7.88E-05	0.00056913
TRIF-mediated TLR3/TLR4 signaling	97	N	7	27	7.88E-05	0.00056913
PI-3K cascade:FGFR2	23	N	4	34	9.35E-05	0.000662226
RAF-independent MAPK1/3 activation	23	N	4	9	9.35E-05	0.000662226
SHC-mediated cascade:FGFR4	23	N	4	47	9.35E-05	0.000662226
TP53 Regulates Transcription of Cell Death Genes	45	Y	5	205	0.0001114	0.000784552
TP53 Regulates Transcription of Cell Cycle Genes	48	Y	5	207	0.000152152	0.001065064
CREB phosphorylation through the activation of Ras	27	N	4	51	0.000179167	0.00123915
VEGFR2 mediated vascular permeability	27	N	4	22	0.000179167	0.00123915
Antigen activates B Cell Receptor (BCR) leading to generation of second messengers	51	N	5	25	0.000203422	0.001398523
Activated TLR4 signalling	113	N	7	27	0.000205497	0.00140443
RIG-I/MDA5 mediated induction of IFN-alpha/beta pathways	80	N	6	18	0.000208714	0.001409737
Senescence-Associated Secretory Phenotype (SASP)	80	N	6	113	0.000208714	0.001409737
Toll-Like Receptors Cascades	151	N	8	28	0.000214461	0.001440129
Downregulation of ERBB2 signaling	29	N	4	87	0.000238479	0.001592154

TFAP2 (AP-2) family regulates transcription of growth factors and their receptors	12	N	3	77	0.0002421 6	0.001607 439
Stabilization of p53	55	Y	5	213	0.0002910 81	0.001921 138
Disassembly of the destruction complex and recruitment of AXIN to the membrane	31	N	4	34	0.0003107 32	0.001982 202
Downregulation of ERBB2:ERBB3 signaling	13	N	3	59	0.0003123 47	0.001982 202
Negative regulation of FGFR4 signaling	31	N	4	26	0.0003107 32	0.001982 202
NICD traffics to nucleus	13	N	3	25	0.0003123 47	0.001982 202
NOD1/2 Signaling Pathway	31	N	4	15	0.0003107 32	0.001982 202
Notch-HLH transcription pathway	13	N	3	25	0.0003123 47	0.001982 202
p38MAPK events	13	N	3	46	0.0003123 47	0.001982 202
TP53 Regulates Metabolic Genes	88	Y	6	204	0.0003508 94	0.002214 661
AMER1 mutants destabilize the destruction complex	14	N	3	27	0.0003944 27	0.002372 724
APC truncation mutants have impaired AXIN binding	14	N	3	27	0.0003944 27	0.002372 724
AXIN missense mutants destabilize the destruction complex	14	N	3	27	0.0003944 27	0.002372 724
AXIN mutants destabilize the destruction complex, activating WNT signaling	14	N	3	27	0.0003944 27	0.002372 724
Regulation of IFNG signaling	14	N	3	8	0.0003944 27	0.002372 724
Regulation of TP53 Activity through Association with Co-factors	14	Y	3	199	0.0003944 27	0.002372 724
TNFR1-induced proapoptotic signaling	14	N	3	6	0.0003944 27	0.002372 724
truncated APC mutants destabilize the destruction complex	14	N	3	27	0.0003944 27	0.002372 724
Truncations of AMER1 destabilize the destruction complex	14	N	3	27	0.0003944 27	0.002372 724
Regulation of TP53 Activity through	91	Y	6	214	0.0004203 94	0.002515 83

Phosphorylation						
RHO GTPase Effectors	257	N	10	34	0.000439659	0.00261756
Prolactin receptor signaling	15	N	3	10	0.000489184	0.002897473
Post NMDA receptor activation events	35	N	4	51	0.000500212	0.002947676
Ovarian tumor domain proteases	36	Y	4	204	0.000558095	0.003255555
Regulation of TNFR1 signaling	36	N	4	15	0.000558095	0.003255555
Regulation of necroptotic cell death	16	N	3	14	0.000597374	0.003467168
TP53 Regulates Transcription of DNA Repair Genes	65	Y	5	203	0.00063437	0.003663486
Apoptotic cleavage of cellular proteins	38	N	4	33	0.000687873	0.003952703
Activation of NMDA receptor upon glutamate binding and postsynaptic events	39	N	4	51	0.000760169	0.004325102
Association of TriC/CCT with target proteins during biosynthesis	39	Y	4	207	0.000760169	0.004325102
Regulation of signaling by CBL	18	N	3	36	0.000856935	0.004828096
RIPK1-mediated regulated necrosis	18	N	3	14	0.000856935	0.004828096
Cyclin A:Cdk2-associated events at S phase entry	70	N	5	105	0.000890603	0.00499343
Intrinsic Pathway for Apoptosis	43	Y	4	200	0.001103453	0.006156949
PI-3K cascade:FGFR4	20	N	3	28	0.001178607	0.006513355
TP53 Regulates Transcription of Genes Involved in Cytochrome C Release	20	Y	3	203	0.001178607	0.006513355
Regulation of gene expression in beta cells	21	N	3	36	0.001364337	0.007468291
TP53 regulates transcription of additional cell cycle genes whose exact role in the p53 pathway remain uncertain	21	Y	3	198	0.001364337	0.007468291
TNF signaling	46	N	4	15	0.001423385	0.007754762

7.4 Gene set enrichment analysis of OAC helpers

Shown are the 189 pathways enriched in drivers (FDR <0.01; 76 also enriched in drivers). For each pathway, the number of genes and samples, the p-value of one-tailed hypergeometric test and the False Discovey Rate (FDR) using the Benjamini and Hochberg method are reported. Pathway size refers to the total number of genes in the pathway. Universal pathways are those with at least one perturbed driver in at least 50% of samples.

Pathway	Pathway size	Universal	Genes (n)	Samples (n)	p-value	FDR
Downstream signal transduction	348	Y	38	84	5.57E-06	0.000146338
NGF signalling via TRKA from the plasma membrane	382	Y	43	93	6.06E-07	2.41E-05
DAP12 signaling	351	Y	37	83	1.63E-05	0.000330733
Fc epsilon receptor (FCERI) signaling	405	Y	44	103	1.19E-06	3.92E-05
DAP12 interactions	366	Y	37	83	4.06E-05	0.000642726
Signaling by the B Cell Receptor (BCR)	242	Y	28	72	3.30E-05	0.000586799
IGF1R signaling cascade	292	Y	33	78	1.11E-05	0.000244331
Insulin receptor signalling cascade	291	Y	33	76	1.04E-05	0.00023934
IRS-mediated signalling	288	Y	32	76	2.12E-05	0.00040842
IRS-related events triggered by IGF1R	292	Y	33	78	1.11E-05	0.000244331
Signaling by Interleukins	402	Y	43	87	2.36E-06	7.56E-05
Downstream signaling events of B Cell Receptor (BCR)	198	Y	26	68	6.70E-06	0.000161252
PI3K/AKT activation	130	Y	17	50	0.000266894	0.002880962
Interleukin-3, 5 and GM-CSF signaling	265	Y	30	66	2.67E-05	0.000488687
PI3K events in ERBB4 signaling	127	Y	17	50	0.000200982	0.002275824
GAB1 signalosome	130	Y	17	50	0.000266894	0.002880962
VEGFA-VEGFR2 Pathway	324	Y	44	128	1.69E-09	4.89E-07
Interleukin-2 signaling	256	Y	28	63	8.97E-05	0.001218676
RET signaling	264	Y	28	60	0.000152142	0.001830459
Interleukin receptor SHC signaling	249	Y	27	62	0.000137433	0.001709468
ARMS-mediated activation	243	Y	24	58	0.001202247	0.008246292
Signalling to p38 via RIT and RIN	243	Y	24	58	0.001202247	0.008246292
Frs2-mediated activation	244	Y	25	60	0.000563708	0.0049041

MAPK1/MAPK3 signaling	245	Y	24	58	0.001345011	0.008928092
Prolonged ERK activation events	246	Y	25	60	0.000635621	0.005281598
Signalling to RAS	250	Y	26	61	0.000350685	0.003432548
Signalling to ERKs	257	Y	27	63	0.000231249	0.002593131
GRB2 events in EGFR signaling	239	Y	24	58	0.000955462	0.0071464 7
RAF/MAP kinase cascade	239	Y	24	58	0.000955462	0.0071464 7
SHC1 events in EGFR signaling	239	Y	24	58	0.000955462	0.0071464 7
SHC1 events in ERBB4 signaling	239	Y	24	58	0.000955462	0.0071464 7
SOS-mediated signalling	239	Y	24	58	0.000955462	0.0071464 7
NCAM signaling for neurite out-growth	270	Y	29	66	9.36E-05	0.001242006
VEGFR2 mediated cell proliferation	252	Y	27	79	0.000167632	0.001975661
Transcriptional Regulation by TP53	366	Y	47	127	3.01E-09	6.94E-07
Signaling by FGFR in disease	63	N	12	26	5.81E-05	0.000860286
Signaling by FGFR1 in disease	38	N	9	24	8.34E-05	0.001146384
Regulation of TP53 Degradation	35	Y	8	9	0.000269524	0.002882405
Regulation of TP53 Expression and Degradation	36	Y	8	9	0.000331475	0.003272256
Cyclin D associated events in G1	38	Y	10	50	1.24E-05	0.000260054
G1 Phase	38	Y	10	50	1.24E-05	0.000260054
PI3K Cascade	79	N	12	29	0.000534811	0.004715315
Mitotic G1-G1/S phases	141	Y	27	102	1.55E-09	4.89E-07
Oxidative Stress Induced Senescence	95	Y	14	49	0.000262466	0.002880962
Signaling by FGFR2	75	N	11	15	0.001213419	0.008246292
Signaling by cytosolic FGFR1 fusion mutants	18	N	5	16	0.001544444	0.009487805
FGFR1 mutant receptor activation	31	N	7	23	0.000704 74	0.005814103
Deubiquitination	280	Y	44	78	1.38E-11	1.59E-08
Oncogene Induced Senescence	34	Y	11	49	4.63E-07	1.98E-05
Regulation of TP53 Activity	159	Y	26	60	9.23E-08	4.84E-06
MET activates RAS signaling	11	N	4	10	0.001552 55	0.009487805
Cyclin E associated events during G1/S transition	71	N	12	24	0.000192947	0.002206469
Signaling by	72	N	11	24	0.000857	0.0066952

NOTCH1					923	79
Ub-specific processing proteases	205	Y	32	60	9.99E-09	1.26E-06
CD28 co-stimulation	33	N	7	25	0.001049226	0.007527058
Pre-NOTCH Expression and Processing	57	Y	10	59	0.000477328	0.004341053
G1/S Transition	118	N	18	88	2.29E-05	0.000433437
Toll Like Receptor 4 (TLR4) Cascade	124	N	23	45	4.67E-08	3.23E-06
C-type lectin receptors (CLRs)	125	N	15	44	0.001468382	0.009487805
p53-Dependent G1 DNA Damage Response	64	Y	10	21	0.00122088	0.008246292
p53-Dependent G1/S DNA damage checkpoint	64	Y	10	21	0.00122088	0.008246292
Signal transduction by L1	21	N	7	25	4.83E-05	0.000724838
G1/S DNA Damage Checkpoints	66	Y	11	22	0.000400571	0.003855496
MyD88-independent TLR3/TLR4 cascade	97	N	20	41	5.59E-08	3.23E-06
Toll Like Receptor 3 (TLR3) Cascade	97	N	20	41	5.59E-08	3.23E-06
TRIF-mediated TLR3/TLR4 signaling	97	N	20	41	5.59E-08	3.23E-06
TP53 Regulates Transcription of Cell Cycle Genes	48	Y	12	52	3.03E-06	8.14E-05
VEGFR2 mediated vascular permeability	27	N	8	35	3.63E-05	0.000599108
Activated TLR4 signalling	113	N	23	45	7.43E-09	1.16E-06
Senescence-Associated Secretory Phenotype (SASP)	80	N	14	29	3.88E-05	0.000623151
Toll-Like Receptors Cascades	151	N	26	53	3.13E-08	2.41E-06
NOD1/2 Signaling Pathway	31	N	8	15	0.000107936	0.001409557
Regulation of TP53 Activity through Phosphorylation	91	Y	13	32	0.000584512	0.004964056
RHO GTPase Effectors	257	N	37	93	6.89E-09	1.16E-06
Apoptotic cleavage of cellular proteins	38	N	9	37	8.34E-05	0.001146384
Cyclin A:Cdk2-associated events at S phase entry	70	N	13	31	3.81E-05	0.000620548
TCR signaling	123	N	15	41	0.001245903	0.008366381
Nucleotide-binding domain, leucine rich repeat containing receptor (NLR)	51	N	11	22	3.53E-05	0.000591694

signaling pathways						
Apoptotic execution phase	52	N	11	45	4.27E-05	0.000666546
S Phase	128	N	22	69	3.66E-07	1.69E-05
MAPK6/MAPK4 signaling	94	N	13	68	0.000800019	0.006372567
L1CAM interactions	100	N	16	43	3.53E-05	0.000591694
SMAD2/SMAD3:SMAD4 heterotrimer regulates transcription	32	N	8	26	0.000137646	0.001709468
Semaphorin interactions	67	N	12	44	0.000108615	0.001409557
MET activates RAP1 and RAC1	11	N	5	13	0.000111679	0.001433209
TRAF6 Mediated Induction of proinflammatory cytokines	72	N	16	35	4.03E-07	1.79E-05
TP53 Regulates Transcription of Genes Involved in G1 Cell Cycle Arrest	13	N	6	39	1.96E-05	0.000384544
Interleukin-1 signaling	44	N	8	16	0.001362342	0.008991454
Transcriptional activity of SMAD2/SMAD3:SMAD4 heterotrimer	44	N	12	39	1.10E-06	3.75E-05
MyD88 cascade initiated on plasma membrane	83	N	16	35	3.01E-06	8.14E-05
Toll Like Receptor 10 (TLR10) Cascade	83	N	16	35	3.01E-06	8.14E-05
Toll Like Receptor 5 (TLR5) Cascade	83	N	16	35	3.01E-06	8.14E-05
TRAF6 mediated induction of NFkB and MAP kinases upon TLR7/8 or 9 activation	83	N	16	35	3.01E-06	8.14E-05
MyD88 dependent cascade initiated on endosome	85	N	17	35	8.66E-07	3.17E-05
Toll Like Receptor 7/8 (TLR7/8) Cascade	85	N	17	35	8.66E-07	3.17E-05
Toll Like Receptor 9 (TLR9) Cascade	89	N	18	36	3.48E-07	1.67E-05
MyD88:Mal cascade initiated on plasma membrane	93	N	20	43	2.64E-08	2.18E-06
Toll Like Receptor 2 (TLR2) Cascade	93	N	20	43	2.64E-08	2.18E-06
Toll Like Receptor TLR1:TLR2 Cascade	93	N	20	43	2.64E-08	2.18E-06
Toll Like Receptor TLR6:TLR2 Cascade	93	N	20	43	2.64E-08	2.18E-06
Negative regulation of MET activity	21	N	6	35	0.000439789	0.004063651

RHO GTPases activate PAKs	21	N	6	25	0.000439789	0.004063651
MAP kinase activation in TLR cascade	60	N	13	26	6.60E-06	0.000161252
SCF(Skp2)-mediated degradation of p27/p21	60	N	10	21	0.000727768	0.005961504
HATs acetylate histones	108	N	14	24	0.000984076	0.007285946
EGFR downregulation	25	N	6	17	0.001208999	0.008246292
HDMs demethylate histones	26	N	6	6	0.00150587	0.009487805
Regulation of lipid metabolism by Peroxisome proliferator-activated receptor alpha (PPARalpha)	117	N	15	78	0.000740 66	0.006022 04
MET promotes cell motility	29	N	11	37	7.05E-08	3.88E-06
G2/M Transition	177	N	20	53	0.000573206	0.0049041
Mitotic G2-G2/M phases	179	N	22	75	9.20E-05	0.001235322
Fcgamma receptor (FCGR) dependent phagocytosis	125	N	21	61	9.91E-07	3.47E-05
Signaling by Robo receptor	32	N	8	36	0.000137646	0.001709468
Positive epigenetic regulation of rRNA expression	76	N	12	36	0.000371701	0.003607688
Regulation of insulin secretion	80	N	12	24	0.000601043	0.005067187
RHO GTPases Activate WASPs and WAVEs	36	N	10	28	7.29E-06	0.00017195
G alpha (12/13) signalling events	87	N	16	49	5.72E-06	0.000146892
EPHB-mediated forward signaling	42	N	14	62	8.04E-09	1.16E-06
Activation of gene expression by SREBF (SREBP)	43	N	8	18	0.001164976	0.008204555
G2/M Checkpoints	151	N	17	58	0.001496911	0.009487805
Resolution of Sister Chromatid Cohesion	100	N	13	26	0.001433972	0.009357273
Cohesin Loading onto Chromatin	10	N	4	5	0.001028597	0.007425182
MET receptor recycling	10	N	5	16	6.35E-05	0.000905996
Establishment of Sister Chromatid Cohesion	11	N	4	5	0.00155255	0.009487805
Regulation of cholesterol biosynthesis by SREBP (SREBF)	56	N	12	28	1.68E-05	0.000334315

RHO GTPases Activate Formins	114	N	18	40	1.41E-05	0.000291288
DEx/H-box helicases activate type I IFN and inflammatory cytokines production	13	N	5	12	0.000286007	0.00303062
Ca2+ pathway	61	N	10	32	0.000832065	0.006537653
p75NTR signals via NF-kB	16	N	6	13	8.05E-05	0.001134003
M Phase	273	N	28	48	0.000266548	0.002880962
G1/S-Specific Transcription	18	N	5	37	0.001544444	0.009487805
MET activates PTK2 signaling	18	N	6	28	0.000171199	0.001997322
Rho GTPase cycle	141	N	23	64	5.26E-07	2.17E-05
Beta-catenin independent WNT signaling	145	N	21	55	1.12E-05	0.000244331
RIP-mediated NFkB activation via ZBP1	21	N	6	17	0.000439789	0.004063651
Deadenylation of mRNA	23	N	6	14	0.0007508	0.00602204
Nuclear Events (kinase and transcription factor activation)	24	N	6	12	0.000959 05	0.00714647
TRAF6 mediated NF-kB activation	24	N	6	17	0.000959 05	0.00714647
G0 and Early G1	25	N	11	47	1.09E-08	1.26E-06
Diseases associated with the TLR signaling cascade	26	N	6	15	0.001505 87	0.009487805
TAK1 activates NFkB by phosphorylation and activation of IKKs complex	26	N	9	23	2.75E-06	8.14E-05
ZBP1(DAI) mediated induction of type I IFNs	26	N	6	17	0.001505 87	0.009487805
HIV Infection	238	N	26	39	0.000160864	0.001915438
Fatty acid, triacylglycerol, and ketone body metabolism	239	N	25	95	0.000414163	0.00392097
Separation of Sister Chromatids	164	N	19	40	0.000571889	0.0049041
EPH-Ephrin signaling	95	N	22	98	1.22E-09	4.89E-07
p75 NTR receptor-mediated signalling	96	N	19	36	2.37E-07	1.19E-05
UCH proteinases	96	N	14	29	0.000293376	0.003034657
MAPK targets/ Nuclear events mediated by MAP kinases	30	N	7	13	0.000569978	0.0049041
mRNA Splicing -	174	N	22	37	6.00E-05	0.0008711

Major Pathway						87
Mitotic Anaphase	175	N	20	42	0.00049508	0.004467323
CLEC7A (Dectin-1) signaling	100	N	13	35	0.001433972	0.009357273
Mitotic Metaphase and Anaphase	176	N	20	42	0.000532903	0.004715315
Regulation of actin dynamics for phagocytic cup formation	100	N	16	51	3.53E-05	0.000591694
mRNA Splicing	185	N	22	37	0.000149675	0.001819732
EPH-ephrin mediated repulsion of cells	48	N	12	34	3.03E-06	8.14E-05
Nuclear Receptor transcription pathway	51	N	11	43	3.53E-05	0.000591694
NRAGE signals death through JNK	59	N	11	27	0.000143643	0.001764972
Orc1 removal from chromatin	71	N	13	53	4.46E-05	0.000677547
Switching of origins to a post-replicative state	71	N	13	53	4.46E-05	0.000677547
Removal of licensing factors from origins	73	N	13	53	6.03E-05	0.000871187
Acetylcholine Neurotransmitter Release Cycle	17	N	5	12	0.001162987	0.008204555
activated TAK1 mediates p38 MAPK activation	17	N	5	11	0.001162987	0.008204555
Activation of ATR in response to replication stress	37	N	8	41	0.000404555	0.003861664
Antiviral mechanism by IFN-stimulated genes	75	N	12	25	0.000327693	0.003265279
Assembly of the pre-replicative complex	68	N	13	78	2.77E-05	0.000499125
CDC6 association with the ORC:origin complex	11	N	4	35	0.00155255	0.009487805
Cell death signalling via NRAGE, NRIF and NADE	75	N	11	27	0.001213419	0.008246292
COPII (Coat Protein 2) Mediated Vesicle Transport	68	N	11	27	0.000521981	0.004673553
Defective CFTR causes cystic fibrosis	61	N	10	22	0.000832065	0.006537653
Deregulated CDK5 triggers multiple neurodegenerative pathways in Alzheimerâ€™s disease models	22	N	8	23	6.54E-06	0.000161252
DNA Replication	108	N	17	84	2.53E-05	0.000471321
DNA Replication Pre-Initiation	85	N	13	78	0.000296897	0.003034657

Dopamine Neurotransmitter Release Cycle	23	N	6	13	0.0007508	0.0060220 4
EPHA-mediated growth cone collapse	34	N	7	59	0.001264 893	0.0084448 07
ER to Golgi Anterograde Transport	131	N	17	38	0.000292 686	0.0030346 57
ISG15 antiviral mechanism	75	N	12	25	0.000327 693	0.0032652 79
M/G1 Transition	85	N	13	78	0.000296 897	0.0030346 57
MHC class II antigen presentation	103	N	14	24	0.000610 358	0.0051084 28
Neutrophil degranulation	479	N	53	115	5.31E-08	3.23E-06
Norepinephrine Neurotransmitter Release Cycle	18	N	5	12	0.001544 444	0.0094878 05
Regulation of DNA replication	76	N	16	84	8.78E-07	3.17E-05
Response to elevated platelet cytosolic Ca2+	133	N	16	69	0.001009 091	0.0074235 69
Serotonin Neurotransmitter Release Cycle	18	N	5	12	0.001544 444	0.0094878 05
Signal attenuation	10	N	4	10	0.001028 597	0.0074251 82
Syndecan interactions	20	N	6	16	0.000327 941	0.0032652 79
Synthesis of DNA	100	N	14	53	0.000450 217	0.0041269 87
Transport to the Golgi and subsequent modification	162	N	20	42	0.000177 381	0.0020487 54
Zinc influx into cells by the SLC39 gene family	10	N	4	12	0.001028 597	0.0074251 82

7.5 First- and co-author papers

Paper 1: Original research article

Citation

Thanos Mourikis, Lorena Benedetti, Elizabeth Foxall, Julianne Perner, Matteo Cereda, Jesper Lagergren, Michael Howell, Christopher Yau, Rebecca Fitzgerald, Paola Scaffidi, Francesca D Ciccarelli. (2018) Patient-specific detection of cancer genes reveals recurrently perturbed processes in esophageal adenocarcinoma. bioRxiv 321612; doi: https://doi.org/10.1101/321612

Abstract

The identification of somatic alterations with a cancer promoting role is challenging in highly unstable and heterogeneous cancers, such as esophageal adenocarcinoma (EAC). Here we used a machine learning approach to identify cancer genes in individual patients considering all types of damaging alterations simultaneously (mutations, copy number alterations and structural rearrangements). Analysing 261 EACs from the OCCAMS Consortium, we discovered a large number of novel cancer genes that, together with well-known drivers, help promote cancer. Validation using 107 additional EACs confirmed the robustness of the approach. Unlike known drivers whose alterations recur across patients, the large majority of the newly discovered cancer genes are rare or patient-specific. Despite this, they converge towards perturbing similar biological processes, including cell cycle progression, proteasome activity, intracellular signalling, Toll-like receptor cascade and DNA replication. Recurrence of process perturbation, rather than individual genes, divides EACs into six clusters that differ in their molecular features and suggest patient stratifications for targeted treatments. Experimental validation of selected genes by mimicking the same alterations found in patients leads to cancer-related phenotypes, thus supporting their contribution to disease progression.

Paper 2: Original research article

Citation

Cereda M., Gambardella G., Benedetti L., Iannelli F., Patel D., Basso G., Guerra R. F., Mourikis T. P., Puccio I., Sinha S., Laghi L., Spencer J., Rodriguez-Justo M., Ciccarelli F. D. (2016) Patients with genetically heterogeneous synchronous colorectal cancer carry rare damaging germline mutations in immune-related genes, Nature Communications. Nature Publishing Group, 7(1), p. 12072. doi: 10.1038/ncomms12072.

Abstract

Synchronous colorectal cancers (syCRCs) are physically separated tumours that develop simultaneously. To understand how the genetic and environmental background influences the development of multiple tumours, here we conduct a comparative analysis of 20 syCRCs from 10 patients. We show that syCRCs have independent genetic origins, acquire dissimilar somatic alterations, and have different clone composition. This inter- and intratumour heterogeneity must be considered in the selection of therapy and in the monitoring of resistance. SyCRC patients show a higher occurrence of inherited damaging mutations in immune-related genes compared to patients with solitary colorectal cancer and to healthy individuals from the 1,000 Genomes Project. Moreover, they have a different composition of immune cell populations in tumour and normal mucosa, and transcriptional differences in immune-related biological processes. This suggests an environmental field effect that promotes multiple tumours likely in the background of inflammation.

Paper 3: Forum

Citation

Cereda, M., Mourikis, T. P. and Ciccarelli, F. D. (2016) Genetic Redundancy, Functional Compensation, and Cancer Vulnerability, *Trends in Cancer*, pp. 160–162. doi: 10.1016/j.trecan.2016.03.003.

Abstract

Cancer genomes acquire somatic alterations that largely differ between and within cancer types. Several of these alterations inactivate genes that are normally functional with no deleterious consequences on cancer cells due to genetic redundancy. Here we discuss how this leads to cancer synthetic dependencies that can be exploited in therapy.

Paper 4: Original research article

Citation

An, O. *et al.* (2016) NCG 5.0: Updates of a manually curated repository of cancer genes and associated properties from cancer mutational screenings, *Nucleic Acids Research*, 44(D1), pp. D992–D999. doi: 10.1093/nar/gkv1123.

Abstract

The Network of Cancer Genes (NCG, http://ncg.kcl.ac.uk/) is a manually curated repository of cancer genes derived from the scientific literature. Due to the increasing amount of cancer genomic data, we have introduced a more robust procedure to extract cancer genes from published cancer mutational screenings and two curators independently reviewed each publication. NCG release 5.0 (August 2015) collects 1571 cancer genes from 175 published studies that describe 188 mutational screenings of 13 315 cancer samples from 49 cancer types and 24 primary sites. In addition to collecting cancer genes, NCG also provides information on the experimental validation that supports the role of these genes in cancer and annotates their properties (duplicability, evolutionary origin, expression profile, function and interactions with proteins and miRNAs).

References

Adzhubei, I A, S Schmidt, L Peshkin, V E Ramensky, A Gerasimova, P Bork, A S Kondrashov, and S R Sunyaev. 2010. “A Method and Server for Predicting Damaging Missense Mutations.” *Nat Methods* 7 (4): 248–49. doi:10.1038/nmeth0410-248.

Agrawal, Nishant, Yuchen Jiao, Chetan Bettegowda, Susan M Hutfless, Yuxuan Wang, Stefan David, Yulan Cheng, et al. 2012. “Comparative Genomic Analysis of Esophageal Adenocarcinoma and Squamous Cell Carcinoma.” *Cancer Discov.* 2 (10): 899–905.

Akira, Shizuo, and Kiyoshi Takeda. 2004. “Toll-like Receptor Signalling.” *Nature Reviews Immunology* 4 (7). Nature Publishing Group: 499–511. doi:10.1038/nri1391.

Aleskandarany, Mohammed A, Ola H Negm, Emad A Rakha, Mohamed A H Ahmed, Christopher C Nolan, Graham R Ball, Carlos Caldas, et al. 2012. “TOMM34 Expression in Early Invasive Breast Cancer: A Biomarker Associated with Poor Outcome.” *Breast Cancer Res Treat* 136: 419–27. doi:10.1007/s10549-012-2249-4.

Alexandrov, L B, S Nik-Zainal, D C Wedge, S A Aparicio, S Behjati, A V Biankin, G R Bignell, et al. 2013. “Signatures of Mutational Processes in Human Cancer.” *Nature* 500 (7463): 415–21.

Allum, William H., Sally P. Stenning, John Bancewicz, Peter I. Clark, and Ruth E. Langley. 2009. “Long-Term Results of a Randomized Trial of Surgery With or Without Preoperative Chemotherapy in Esophageal Cancer.” *Journal of Clinical Oncology* 27 (30): 5062–67. doi:10.1200/JCO.2009.22.2083.

An, Omer, Giovanni M. Dall’Olio, Thanos P. Mourikis, and Francesca D. Ciccarelli. 2016. “NCG 5.0: Updates of a Manually Curated Repository of Cancer Genes and Associated Properties from Cancer Mutational Screenings.” *Nucleic Acids Research* 44 (D1): D992–99. doi:10.1093/nar/gkv1123.

Anaparthy, Rajeswari, and Prateek Sharma. 2014. “Progression of Barrett Oesophagus: Role of Endoscopic and Histological Predictors.” *Nature Reviews Gastroenterology & Hepatology* 11 (9): 525–34. doi:10.1038/nrgastro.2014.69.

Anthony, Gidudu, and Heinz Ruther. 2007. “Comparison of Feature Selection Techniques for SVM Classification.” In *In 10th International Symposium on Physical Measurements and Signatures in Remote Sensing*. https://pdfs.semanticscholar.org/5a87/cce0f52e7f9ca8ee19058a06d2f0bcbaca88.pdf.

Armitage, P, and R Doll. 1954. “The Age Distribution of Cancer and a Multi-Stage Theory of Carcinogenesis.” *British Journal of Cancer* 8 (1). Nature Publishing Group: 1–12. http://www.ncbi.nlm.nih.gov/pubmed/13172380.

Bailey, Matthew H., Collin Tokheim, Eduard Porta-Pardo, Sohini Sengupta, Denis Bertrand, Amila Weerasinghe, Antonio Colaprico, et al. 2018. “Comprehensive Characterization of Cancer Driver Genes and Mutations.” *Cell* 173 (2). Cell Press: 371–385.e18. doi:10.1016/J.CELL.2018.02.060.

Bakhoum, S F, B Ngo, A M Laughney, J A Cavallo, C J Murphy, P Ly, P Shah, et al. 2018. “Chromosomal Instability Drives Metastasis through a Cytosolic DNA Response.” *Nature* 553 (7689): 467–72. doi:10.1038/nature25432.

Bang, Y J, E Van Cutsem, A Feyereislova, H C Chung, L Shen, A Sawaki, F Lordick, et al. 2010. “Trastuzumab in Combination with Chemotherapy versus Chemotherapy Alone for Treatment of HER2-Positive Advanced Gastric or Gastro-Oesophageal Junction Cancer (ToGA): A Phase 3, Open-Label, Randomised Controlled Trial.” *Lancet* 376 (9742): 687–97. doi:10.1016/S0140-6736(10)61121-X.

Ben-Aroya, Shay, and Erez Y. Levanon. 2018. “A-to-I RNA Editing: An Overlooked Source of Cancer Mutations.” *Cancer Cell* 33 (5). Cell Press: 789–90. doi:10.1016/J.CCELL.2018.04.006.

Benedetti, Lorena, Matteo Cereda, LeeAnn Monteverde, Nikita Desai, Francesca D. Ciccarelli, Lorena Benedetti, Matteo Cereda, LeeAnn Monteverde, Nikita Desai, and Francesca D. Ciccarelli. 2017. “Synthetic Lethal Interaction between the Tumour Suppressor STAG2 and Its Paralog STAG1.” *Oncotarget* 8 (23). Impact Journals: 37619–32. doi:10.18632/oncotarget.16838.

Bennett, Laura, Matthew Howell, Danish Memon, Chris Smowton, Cong Zhou, and Crispin J. Miller. 2018. “Mutation Pattern Analysis Reveals Polygenic Mini-Drivers Associated with Relapse after Surgery in Lung Adenocarcinoma.” *Scientific Reports* 8 (1). Nature Publishing Group: 14830. doi:10.1038/s41598-018-33276-3.

Bishop, Chris M. 1994. “Novelty Detection and Neural Network Validation.” *IEE Proceedings on Vision, Image and Signal Processing. Special Issue on Applications of Neural Networks* 141: 217–22. https://pdfs.semanticscholar.org/7e7d/844569a5cff8b568d3d291aa99dd11dbfa23.pdf.

Blanke, Charles D., Cathryn Rankin, George D. Demetri, Christopher W. Ryan, Margaret von Mehren, Robert S. Benjamin, A. Kevin Raymond, et al. 2008. “Phase III Randomized, Intergroup Trial Assessing Imatinib Mesylate At Two Dose Levels in Patients With Unresectable or Metastatic Gastrointestinal Stromal Tumors Expressing the Kit Receptor Tyrosine Kinase: S0033.” *Journal of Clinical Oncology* 26 (4): 626–32. doi:10.1200/JCO.2007.13.4452.

Blum, Avrim, and Tom Mitchell. 1998. “Combining Labeled and Unlabeled Data with Co-Training.” In *COLT’ 98 Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, 92–100. http://www.cs.cmu.edu/~avrim/Papers/cotrain.pdf.

Bochman, Matthew L, and Anthony Schwacha. 2009. “The Mcm Complex: Unwinding the Mechanism of a Replicative Helicase.” *Microbiology and Molecular Biology Reviews : MMBR* 73 (4). American Society for Microbiology (ASM): 652–83. doi:10.1128/MMBR.00019-09.

Boveri, T. 1914. *Zur Frage Der Entstehung Maligner Tumoren.* Gustav Fisher.

Breunig, Markus M, Hans-Peter Kriegel, Raymond T Ng, and Jörg Sander. 2000. “LOF: Identifying Density-Based Local Outliers.” In *Proc. ACM SIGMOD 2000 Int. Conf. On Management of Data, Dalles, TX, 2000*. http://www.dbs.ifi.lmu.de/Publikationen/Papers/LOF.pdf.

Brighton, Henry, and Gerd Gigerenzer. 2015. “The Bias Bias.” *Journal of Business Research* 68 (8). Elsevier: 1772–84. doi:10.1016/J.JBUSRES.2015.01.061.

Brown, K. R., and I. Jurisica. 2005. “Online Predicted Human Interaction Database.” *Bioinformatics* 21 (9). Oxford University Press: 2076–82. doi:10.1093/bioinformatics/bti273.

Cabestany, J, A Prieto, D F Sandoval, Michel Verleysen, and Damien François. 2005. “The Curse of Dimensionality in Data Mining and Time Series Prediction.” *LNCS*. Vol. 3512. www.ucl.ac.be/mlg.

Cahill, D P, K W Kinzler, B Vogelstein, and C Lengauer. 1999. “Genetic

Instability and Darwinian Selection in Tumours." *Trends in Cell Biology* 9 (12): M57-60. http://www.ncbi.nlm.nih.gov/pubmed/10611684.

Calvano, Steve E., Wenzhong Xiao, Daniel R. Richards, Ramon M. Felciano, Henry V. Baker, Raymond J. Cho, Richard O. Chen, et al. 2005. "A Network-Based Analysis of Systemic Inflammation in Humans." *Nature* 437 (7061). Nature Publishing Group: 1032–37. doi:10.1038/nature03985.

Cancer Genome Atlas Research Network. 2014. "Comprehensive Molecular Profiling of Lung Adenocarcinoma." *Nature* 511 (7511): 543–50. doi:10.1038/nature13385.

Carter, S L, A C Eklund, I S Kohane, L N Harris, and Z Szallasi. 2006. "A Signature of Chromosomal Instability Inferred from Gene Expression Profiles Predicts Clinical Outcome in Multiple Human Cancers." *Nat Genet* 38 (9): 1043–48. doi:10.1038/ng1861.

Castro-Giner, F, P Ratcliffe, and I Tomlinson. 2015. "The Mini-Driver Model of Polygenic Cancer Evolution." *Nat Rev Cancer* 15 (11): 680–85. doi:10.1038/nrc3999.

Chan, Timothy A, Mark Yarchoan, Elizabeth Jaffee, Charles Swanton, Sergio A Quezada, Albrecht Stenzinger, and Solange Peters. 2018. "Development of Tumor Mutation Burden as an Immunotherapy Biomarker: Utility for the Oncology Clinic." *Annals of Oncology*, November. doi:10.1093/annonc/mdy495.

Chandar, Apoorva Krishna, Swapna Devanna, Chang Lu, Siddharth Singh, Katarina Greer, Amitabh Chak, and Prasad G Iyer. 2015. "Association of Serum Levels of Adipokines and Insulin With Risk of Barrett's Esophagus: A Systematic Review and Meta-Analysis." *Clinical Gastroenterology and Hepatology : The Official Clinical Practice Journal of the American Gastroenterological Association* 13 (13): 2241–55.e1–4; quiz e179. doi:10.1016/j.cgh.2015.06.041.

Chapelle, Olivier, and Olivier Bousquet. 2002. "Choosing Multiple Parameters for Support Vector Machines." *Machine Learning* 46: 131–59. https://link.springer.com/content/pdf/10.1023/A:1012450327387.pdf.

Chatr-aryamontri, Andrew, Rose Oughtred, Lorrie Boucher, Jennifer Rust, Christie Chang, Nadine K. Kolas, Lara O'Donnell, et al. 2017. "The BioGRID Interaction Database: 2017 Update." *Nucleic Acids Research* 45 (D1): D369–79. doi:10.1093/nar/gkw1102.

Chatterjee, Aniruddha, Euan J. Rodger, and Michael R. Eccles. 2018. “Epigenetic Drivers of Tumourigenesis and Cancer Metastasis.” *Seminars in Cancer Biology* 51 (August). Academic Press: 149–59. doi:10.1016/J.SEMCANCER.2017.08.004.

Chawla, Nitesh V, and Grigoris Karakoulas. 2005. “Learning From Labeled And Unlabeled Data: An Empirical Study Across Techniques And Domains.” *Journal of Artificial Intelligence Research*. Vol. 23. https://arxiv.org/pdf/1109.2047.pdf.

Chen, Long, Tingyi Wei, Xiaoxing Si, Qianqian Wang, Yan Li, Ye Leng, Anmei Deng, et al. 2013. “GCN5 Potentiates Lung Cancer Growth Via E2F1 Lysine Acetyltransferase GCN5 Potentiates the Growth of Non-Small Cell Lung Cancer via Promotion of E2F1, Cyclin D1 and Cyclin E1 Expression.” doi:10.1074/jbc.M113.458737.

Chen, X, O Schulz-Trieglaff, R Shaw, B Barnes, F Schlesinger, M Kallberg, A J Cox, S Kruglyak, and C T Saunders. 2016. “Manta: Rapid Detection of Structural Variants and Indels for Germline and Cancer Sequencing Applications.” *Bioinformatics* 32 (8): 1220–22.

Chen, Zezhi, Nick Pears, Michael Freeman, and Jim Austin. 2009. “Road Vehicle Classification Using Support Vector Machines.” In *2009 IEEE International Conference on Intelligent Computing and Intelligent Systems*, 214–18. IEEE. doi:10.1109/ICICISYS.2009.5357707.

Chin, Lynda, William C Hahn, Gad Getz, and Matthew Meyerson. 2011. “Making Sense of Cancer Genomic Data.” *Genes & Development* 25 (6). Cold Spring Harbor Laboratory Press: 534–55. doi:10.1101/gad.2017311.

Chow, W H, W D Finkle, J K McLaughlin, H Frankl, H K Ziel, and J F Fraumeni. 1995. “The Relation of Gastroesophageal Reflux Disease and Its Treatment to Adenocarcinomas of the Esophagus and Gastric Cardia.” *JAMA* 274 (6): 474–77. http://www.ncbi.nlm.nih.gov/pubmed/7629956.

Chun, S, and J C Fay. 2009. “Identification of Deleterious Mutations within Three Human Genomes.” *Genome Res.* 19 (9): 1553–61.

Ciccarelli, F D. 2010. “The (r)Evolution of Cancer Genetics.” *BMC Biol.* 8: 74.

Ciriello, G, E Cerami, C Sander, and N Schultz. 2012. “Mutual Exclusivity Analysis Identifies Oncogenic Network Modules.” *Genome Res* 22 (2): 398–406. doi:10.1101/gr.125567.111.

Coleman, Helen G., Shao-Hua Xie, and Jesper Lagergren. 2018. “The

Epidemiology of Esophageal Adenocarcinoma." *Gastroenterology* 154 (2). W.B. Saunders: 390–405. doi:10.1053/J.GASTRO.2017.07.046.

Collins, Francis S, and Anna D Barker. 2007. "Mapping the Cancer Genome. Pinpointing the Genes Involved in Cancer Will Help Chart a New Course across the Complex Landscape of Human Malignancies." *Scientific American* 296 (3): 50–57. http://www.ncbi.nlm.nih.gov/pubmed/17348159.

Cook, D R, K L Rossman, and C J Der. 2014. "Rho Guanine Nucleotide Exchange Factors: Regulators of Rho GTPase Activity in Development and Disease." *Oncogene* 33 (31): 4021–35. doi:10.1038/onc.2013.362.

Cook, J. W., C. L. Hewett, and I. Hieger. 1933. "The Isolation of a Cancer-Producing Hydrocarbon from Coal Tar. Parts I, II, and III." *Journal of the Chemical Society* 0 (0). The Royal Society of Chemistry: 395. doi:10.1039/jr9330000395.

Cook, M B, W-H Chow, and S S Devesa. 2009. "Oesophageal Cancer Incidence in the United States by Race, Sex, and Histologic Type, 1977-2005." *British Journal of Cancer* 101 (5). Nature Publishing Group: 855–59. doi:10.1038/sj.bjc.6605246.

Cook, Michael B., Farin Kamangar, David C. Whiteman, Neal D. Freedman, Marilie D. Gammon, Leslie Bernstein, Linda M. Brown, et al. 2010. "Cigarette Smoking and Adenocarcinomas of the Esophagus and Esophagogastric Junction: A Pooled Analysis From the International BEACON Consortium." *JNCI: Journal of the National Cancer Institute* 102 (17): 1344–53. doi:10.1093/jnci/djq289.

Cook, Michael B., Nicholas J. Shaheen, Lesley A. Anderson, Carol Giffen, Wong–Ho Chow, Thomas L. Vaughan, David C. Whiteman, and Douglas A. Corley. 2012. "Cigarette Smoking Increases Risk of Barrett's Esophagus: An Analysis of the Barrett's and Esophageal Adenocarcinoma Consortium." *Gastroenterology* 142 (4): 744–53. doi:10.1053/j.gastro.2011.12.049.

Cook, Michael B, Douglas A Corley, Liam J Murray, Linda M Liao, Farin Kamangar, Weimin Ye, Marilie D Gammon, et al. 2014. "Gastroesophageal Reflux in Relation to Adenocarcinomas of the Esophagus: A Pooled Analysis from the Barrett's and Esophageal Adenocarcinoma Consortium (BEACON)." *PloS One* 9 (7). Public Library of Science: e103508. doi:10.1371/journal.pone.0103508.

Cook, Michael Blaise. 2011. "Editorial: Non-Acid Reflux: The Missing Link

Between Gastric Atrophy and Esophageal Squamous Cell Carcinoma?" *The American Journal of Gastroenterology* 106 (11): 1930–32. doi:10.1038/ajg.2011.288.

Cortes, Corinna, and Vladimir Vapnik. 1995. "Support-Vector Networks." *Machine Learning* 20 (3). Kluwer Academic Publishers: 273–97. doi:10.1007/BF00994018.

Crawford, Sarah. 2013. "Is It Time for a New Paradigm for Systemic Cancer Treatment? Lessons from a Century of Cancer Chemotherapy." *Frontiers in Pharmacology* 4. Frontiers Media SA: 68. doi:10.3389/fphar.2013.00068.

Cunningham, David, William H. Allum, Sally P. Stenning, Jeremy N. Thompson, Cornelis J.H. Van de Velde, Marianne Nicolson, J. Howard Scarffe, et al. 2006. "Perioperative Chemotherapy versus Surgery Alone for Resectable Gastroesophageal Cancer." *New England Journal of Medicine* 355 (1): 11–20. doi:10.1056/NEJMoa055531.

Cutsem, Eric Van, Claus-Henning Köhne, Erika Hitre, Jerzy Zaluski, Chung-Rong Chang Chien, Anatoly Makhson, Geert D'Haens, et al. 2009. "Cetuximab and Chemotherapy as Initial Treatment for Metastatic Colorectal Cancer." *New England Journal of Medicine* 360 (14): 1408–17. doi:10.1056/NEJMoa0805019.

Cutsem, Eric Van, Claus-Henning Köhne, István Láng, Gunnar Folprecht, Marek P. Nowacki, Stefano Cascinu, Igor Shchepotin, et al. 2011. "Cetuximab Plus Irinotecan, Fluorouracil, and Leucovorin As First-Line Treatment for Metastatic Colorectal Cancer: Updated Analysis of Overall Survival According to Tumor *KRAS* and *BRAF* Mutation Status." *Journal of Clinical Oncology* 29 (15): 2011–19. doi:10.1200/JCO.2010.33.5091.

D'Antonio, M, and F D Ciccarelli. 2011. "Modification of Gene Duplicability during the Evolution of Protein Interaction Network." *PLoS Comput Biol* 7 (4): e1002029. doi:10.1371/journal.pcbi.1002029.

D'Antonio, M, V Pendino, S Sinha, and F D Ciccarelli. 2012. "Network of Cancer Genes (NCG 3.0): Integration and Analysis of Genetic and Network Properties of Cancer Genes." *Nucleic Acids Res* 40 (Database issue): D978-83. doi:10.1093/nar/gkr952.

D'Antonio, Matteo, and Francesca D Ciccarelli. 2013. "Integrated Analysis of Recurrent Properties of Cancer Genes to Identify Novel Drivers." *Genome Biology* 14 (5): R52. doi:10.1186/gb-2013-14-5-r52.

Dang, Chi V. 2012. “MYC on the Path to Cancer.” *Cell* 149 (1): 22–35. doi:10.1016/j.cell.2012.03.003.

Davydov, E V, D L Goode, M Sirota, G M Cooper, A Sidow, and S Batzoglou. 2010. “Identifying a High Fraction of the Human Genome to Be under Selective Constraint Using GERP++.” *PLoS Comput Biol* 6 (12): e1001025. doi:10.1371/journal.pcbi.1001025.

Dawsey, S M, K J Lewin, F S Liu, G Q Wang, and Q Shen. 1994. “Esophageal Morphology from Linxian, China. Squamous Histologic Findings in 754 Patients.” *Cancer* 73 (8): 2027–37.

Dawsey, S M, K J Lewin, G Q Wang, F S Liu, R K Nieberg, Y Yu, J Y Li, W J Blot, B Li, and P R Taylor. 1994. “Squamous Esophageal Histology and Subsequent Risk of Squamous Cell Carcinoma of the Esophagus. A Prospective Follow-up Study from Linxian, China.” *Cancer* 74 (6): 1686–92. http://www.ncbi.nlm.nih.gov/pubmed/8082069.

Dees, N D, Q Zhang, C Kandoth, M C Wendl, W Schierding, D C Koboldt, T B Mooney, et al. 2012. “MuSiC: Identifying Mutational Significance in Cancer Genomes.” *Genome Res* 22 (8): 1589–98. doi:10.1101/gr.134635.111.

Deininger, M. W. N., and Brian J Druker. 2003. “Specific Targeted Therapy of Chronic Myelogenous Leukemia with Imatinib.” *Pharmacological Reviews* 55 (3): 401–23. doi:10.1124/pr.55.3.4.

Ding, L, M C Wendl, J F McMichael, and B J Raphael. 2014. “Expanding the Computational Toolbox for Mining Cancer Genomes.” *Nat Rev Genet* 15 (8): 556–70. doi:10.1038/nrg3767.

Domazet-Lošo, Tomislav, and Diethard Tautz. 2010. “Phylostratigraphic Tracking of Cancer Genes Suggests a Link to the Emergence of Multicellularity in Metazoa.” *BMC Biology* 8 (1): 66. doi:10.1186/1741-7007-8-66.

Douillard, Jean-Yves, Frances A. Shepherd, Vera Hirsh, Tony Mok, Mark A. Socinski, Radj Gervais, Mei-Lin Liao, et al. 2010. “Molecular Predictors of Outcome With Gefitinib and Docetaxel in Previously Treated Non–Small-Cell Lung Cancer: Data From the Randomized Phase III INTEREST Trial.” *Journal of Clinical Oncology* 28 (5): 744–52. doi:10.1200/JCO.2009.24.3030.

Druker, Brian J. 2004. “Imatinib as a Paradigm of Targeted Therapies.” In *Advances in Cancer Research*, 91:1–30. doi:10.1016/S0065-

230X(04)91001-9.

Druker, Brian J., Charles L. Sawyers, Hagop Kantarjian, Debra J. Resta, Sofia Fernandes Reese, John M. Ford, Renaud Capdeville, and Moshe Talpaz. 2001. “Activity of a Specific Inhibitor of the BCR-ABL Tyrosine Kinase in the Blast Crisis of Chronic Myeloid Leukemia and Acute Lymphoblastic Leukemia with the Philadelphia Chromosome.” *New England Journal of Medicine* 344 (14): 1038–42. doi:10.1056/NEJM200104053441402.

Druker, Brian J., Moshe Talpaz, Debra J. Resta, Bin Peng, Elisabeth Buchdunger, John M. Ford, Nicholas B. Lydon, et al. 2001. “Efficacy and Safety of a Specific Inhibitor of the BCR-ABL Tyrosine Kinase in Chronic Myeloid Leukemia.” *New England Journal of Medicine* 344 (14): 1031–37. doi:10.1056/NEJM200104053441401.

Druker, Brian J. 2003. “Imatinib Mesylate in the Treatment of Chronic Myeloid Leukaemia.” *Expert Opinion on Pharmacotherapy* 4 (6): 963–71. doi:10.1517/14656566.4.6.963.

Dulak, A M, P Stojanov, S Peng, M S Lawrence, C Fox, C Stewart, S Bandla, et al. 2013. “Exome and Whole-Genome Sequencing of Esophageal Adenocarcinoma Identifies Recurrent Driver Events and Mutational Complexity.” *Nat Genet* 45 (5): 478–86. doi:10.1038/ng.2591.

Ek, W. E., D. M. Levine, M. D’Amato, N. L. Pedersen, P. K. E. Magnusson, F. Bresso, L. E. Onstad, et al. 2013. “Germline Genetic Contributions to Risk for Esophageal Adenocarcinoma, Barrett’s Esophagus, and Gastroesophageal Reflux.” *JNCI Journal of the National Cancer Institute* 105 (22). Oxford University Press: 1711–18. doi:10.1093/jnci/djt303.

Emens, Leisha A., Paolo A. Ascierto, Phillip K. Darcy, Sandra Demaria, Alexander M.M. Eggermont, William L. Redmond, Barbara Seliger, and Francesco M. Marincola. 2017. “Cancer Immunotherapy: Opportunities and Challenges in the Rapidly Evolving Clinical Landscape.” *European Journal of Cancer* 81 (August). Pergamon: 116–29. doi:10.1016/J.EJCA.2017.01.035.

Enders, Craig K. 2010. *Applied Missing Data Analysis*. Guilford Press. https://www.guilford.com/books/Applied-Missing-Data-Analysis/Craig-Enders/9781606236390.

Espinosa, Maria Claudia, Muhammad Attiq Rehman, Patricia Chisamore-Robert, Daniel Jeffery, and Krassimir Yankulov. 2010. “GCN5 Is a Positive

Regulator of Origins of DNA Replication in Saccharomyces Cerevisiae." Edited by Anja-Katrin Bielinsky. *PLoS ONE* 5 (1): e8964. doi:10.1371/journal.pone.0008964.

Fabregat, A, K Sidiropoulos, P Garapati, M Gillespie, K Hausmann, R Haw, B Jassal, et al. 2016. "The Reactome Pathway Knowledgebase." *Nucleic Acids Res.* 44 (D1): D481--7.

Farria, A, W Li, and S Y R Dent. 2015. "KATs in Cancer: Functions and Therapies." *Oncogene* 34 (38). NIH Public Access: 4901–13. doi:10.1038/onc.2014.453.

Farrow, D C, T L Vaughan, C Sweeney, M D Gammon, W H Chow, H A Risch, J L Stanford, et al. 2000. "Gastroesophageal Reflux Disease, Use of H2 Receptor Antagonists, and Risk of Esophageal and Gastric Cancer." *Cancer Causes & Control : CCC* 11 (3): 231–38. http://www.ncbi.nlm.nih.gov/pubmed/10782657.

Fawcett, Tom. 2006. "An Introduction to ROC Analysis." *Pattern Recognition Letters* 27 (8). North-Holland: 861–74. doi:10.1016/J.PATREC.2005.10.010.

Fei, Liangru, and Hongtao Xu. 2018. "Role of MCM2-7 Protein Phosphorylation in Human Cancer Cells." *Cell & Bioscience* 8. BioMed Central: 43. doi:10.1186/s13578-018-0242-2.

Fels Elliott, D R, J Perner, X Li, M F Symmons, B Verstak, M Eldridge, L Bower, et al. 2017. "Impact of Mutations in Toll-like Receptor Pathway Genes on Esophageal Carcinogenesis." *PLoS Genet.* 13 (5): e1006808.

Ferlay, Jacques, Isabelle Soerjomataram, Rajesh Dikshit, Sultan Eser, Colin Mathers, Marise Rebelo, Donald Maxwell Parkin, David Forman, and Freddie Bray. 2015. "Cancer Incidence and Mortality Worldwide: Sources, Methods and Major Patterns in GLOBOCAN 2012." *International Journal of Cancer* 136 (5): E359–86. doi:10.1002/ijc.29210.

Forbes, S A, D Beare, P Gunasekaran, K Leung, N Bindal, H Boutselakis, M Ding, et al. 2015. "COSMIC: Exploring the World's Knowledge of Somatic Mutations in Human Cancer." *Nucleic Acids Res.* 43 (Database issue): D805--11.

Forbes, Simon A., David Beare, Harry Boutselakis, Sally Bamford, Nidhi Bindal, John Tate, Charlotte G. Cole, et al. 2017. "COSMIC: Somatic Cancer Genetics at High-Resolution." *Nucleic Acids Research* 45 (D1): D777–83.

doi:10.1093/nar/gkw1121.

Freedman, David. 2009. *Statistical Models : Theory and Practice*. Cambridge University Press.

Friedenberg, Frank K., Melissa Xanthopoulos, Gary D. Foster, and Joel E. Richter. 2008. “The Association Between Gastroesophageal Reflux Disease and Obesity.” *The American Journal of Gastroenterology* 103 (8). Nature Publishing Group: 2111–22. doi:10.1111/j.1572-0241.2008.01946.x.

Futreal, P A, L Coin, M Marshall, T Down, T Hubbard, R Wooster, N Rahman, and M R Stratton. 2004. “A Census of Human Cancer Genes.” *Nat. Rev. Cancer* 4 (3): 177–83.

Galanos, Panagiotis, Konstantinos Vougas, David Walter, Alexander Polyzos, Apolinar Maya-Mendoza, Emma J. Haagensen, Antonis Kokkalis, et al. 2016. “Chronic P53-Independent P21 Expression Causes Genomic Instability by Deregulating Replication Licensing.” *Nature Cell Biology* 18 (7): 777–89. doi:10.1038/ncb3378.

Garassino, M. C., M. Marabese, P. Rusconi, E. Rulli, O. Martelli, G. Farina, A. Scanni, and M. Broggini. 2011. “Different Types of K-Ras Mutations Could Affect Drug Sensitivity and Tumour Behaviour in Non-Small-Cell Lung Cancer.” *Annals of Oncology* 22 (1): 235–37. doi:10.1093/annonc/mdq680.

Garber, M, M Guttman, M Clamp, M C Zody, N Friedman, and X Xie. 2009. “Identifying Novel Constrained Elements by Exploiting Biased Substitution Patterns.” *Bioinformatics* 25 (12): i54-62. doi:10.1093/bioinformatics/btp190.

Garraway, L A, and E S Lander. 2013. “Lessons from the Cancer Genome.” *Cell* 153 (1): 17–37. doi:10.1016/j.cell.2013.03.002.

Geman, Stuart, Elie Bienenstock, and René Doursat. 1992. “Neural Networks and the Bias/Variance Dilemma.” *Neural Computation* 4 (1). MIT Press: 1–58. doi:10.1162/neco.1992.4.1.1.

Getz, G, H Hofling, J P Mesirov, T R Golub, M Meyerson, R Tibshirani, and E S Lander. 2007. “Comment on ‘The Consensus Coding Sequences of Human Breast and Colorectal Cancers.’” *Science* 317 (5844): 1500. doi:10.1126/science.1138764.

Ghajar, Cyrus M, Héctor Peinado, Hidetoshi Mori, Irina R Matei, Kimberley J Evason, Hélène Brazier, Dena Almeida, et al. 2013. “The Perivascular Niche Regulates Breast Tumour Dormancy.” *Nature Cell Biology* 15 (7).

NIH Public Access: 807–17. doi:10.1038/ncb2767.

Giaginis, Constantinos, Stephanie Vgenopoulou, Philippe Vielh, and Stamatios Theocharis. 2010. “MCM Proteins as Diagnostic and Prognostic Tumor Markers in the Clinical Setting.” *Histology and Histopathology* 25 (3): 351–70. doi:10.14670/HH-25.351.

Gillies, Robert J., Paul E. Kinahan, and Hedvig Hricak. 2016. “Radiomics: Images Are More than Pictures, They Are Data.” *Radiology* 278 (2): 563–77. doi:10.1148/radiol.2015151169.

Giot, L, J S Bader, C Brouwer, A Chaudhuri, B Kuang, Y Li, Y L Hao, et al. 2003. “A Protein Interaction Map of Drosophila Melanogaster.” *Science (New York, N.Y.)* 302 (5651). American Association for the Advancement of Science: 1727–36. doi:10.1126/science.1090289.

Golub, T R, D K Slonim, P Tamayo, C Huard, M Gaasenbeek, J P Mesirov, H Coller, et al. 1999. “Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring.” *Science (New York, N.Y.)* 286 (5439): 531–37. http://www.ncbi.nlm.nih.gov/pubmed/10521349.

Gonzalez-Perez, A, and N Lopez-Bigas. 2012. “Functional Impact Bias Reveals Cancer Drivers.” *Nucleic Acids Res* 40 (21): e169. doi:10.1093/nar/gks743.

Gonzalez-Perez, Abel, Christian Perez-Llamas, Jordi Deu-Pons, David Tamborero, Michael P. Schroeder, Alba Jene-Sanz, Alberto Santos, and Nuria Lopez-Bigas. 2013. “IntOGen-Mutations Identifies Cancer Drivers across Tumor Types.” *Nature Methods* 10 (11): 1081–84. doi:10.1038/nmeth.2642.

Govindan, Ramaswamy, Li Ding, Malachi Griffith, Janakiraman Subramanian, Nathan D. Dees, Krishna L. Kanchi, Christopher A. Maher, et al. 2012. “Genomic Landscape of Non-Small Cell Lung Cancer in Smokers and Never-Smokers.” *Cell* 150 (6): 1121–34. doi:10.1016/j.cell.2012.08.024.

Greaves, M, and C C Maley. 2012. “Clonal Evolution in Cancer.” *Nature* 481 (7381): 306–13. doi:10.1038/nature10762.

Gu, Z, R Eils, and M Schlesner. 2016. “Complex Heatmaps Reveal Patterns and Correlations in Multidimensional Genomic Data.” *Bioinformatics* 32 (18): 2847–49.

Guyon, Isabelle, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. 2002. “Gene Selection for Cancer Classification Using Support Vector Machines.” *Machine Learning* 46 (1/3). Kluwer Academic Publishers: 389–422.

doi:10.1023/A:1012487302797.

Haber, Daniel A., and Jeff Settleman. 2007. “Drivers and Passengers.” *Nature* 446 (7132): 145–46. doi:10.1038/446145a.

Hagen, P. van, M.C.C.M. Hulshof, J.J.B. van Lanschot, E.W. Steyerberg, M.I. van Berge Henegouwen, B.P.L. Wijnhoven, D.J. Richel, et al. 2012. “Preoperative Chemoradiotherapy for Esophageal or Junctional Cancer.” *New England Journal of Medicine* 366 (22): 2074–84. doi:10.1056/NEJMoa1112088.

Han, Leng, Lixia Diao, Shuangxing Yu, Xiaoyan Xu, Jie Li, Rui Zhang, Yang Yang, et al. 2015. “The Genomic Landscape and Clinical Relevance of A-to-I RNA Editing in Human Cancers.” *Cancer Cell* 28 (4). Cell Press: 515–28. doi:10.1016/J.CCELL.2015.08.013.

Hanahan, D, and R A Weinberg. 2000. “The Hallmarks of Cancer.” *Cell* 100 (1): 57–70. https://www.ncbi.nlm.nih.gov/pubmed/10647931.

Hanahan, D, and R A Weinberg.. 2011. “Hallmarks of Cancer: The next Generation.” *Cell* 144 (5): 646–74. doi:10.1016/j.cell.2011.02.013.

Hansemann, D von. 1890. “Ueber Asymmetrische Zelltheilung in Epithel Krebsen Und Biologische Bedeutung.” *Virchows Arch. Path. Anat.*, no. 119: 299.

Hao, Yujun, Chao Wang, Bo Cao, Brett M. Hirsch, Jing Song, Sanford D. Markowitz, Rob M. Ewing, et al. 2013. “Gain of Interaction with IRS1 by P110α-Helical Domain Mutants Is Crucial for Their Oncogenic Functions.” *Cancer Cell* 23 (5): 583–93. doi:10.1016/j.ccr.2013.03.021.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. “Unsupervised Learning.” In *The Elements of Statistical Learning*, 485–585. Springer, New York, NY. doi:10.1007/978-0-387-84858-7_14.

Hiley, C, E C de Bruin, N McGranahan, and C Swanton. 2014. “Deciphering Intratumor Heterogeneity and Temporal Acquisition of Driver Events to Refine Precision Medicine.” *Genome Biol* 15 (8): 453. doi:10.1186/s13059-014-0453-8.

Hills, S A, and J F X Diffley. 2014. “{DNA} Replication and {Oncogene-Induced} Replicative Stress (Vol 24, Pg R435, 2014).” *Curr. Biol.* 24 (13): 1563.

Hoadley, Katherine A, Christina Yau, Denise M Wolf, Andrew D Cherniack, David Tamborero, Sam Ng, Max D M Leiserson, et al. 2014. “Multiplatform Analysis of 12 Cancer Types Reveals Molecular Classification within and

across Tissues of Origin." *Cell* 158 (4). NIH Public Access: 929–44. doi:10.1016/j.cell.2014.06.049.

Hochhaus, A, S G O'Brien, F Guilhot, B J Druker, S Branford, L Foroni, J M Goldman, et al. 2009. "Six-Year Follow-up of Patients Receiving Imatinib for the First-Line Treatment of Chronic Myeloid Leukemia." *Leukemia* 23 (6): 1054–61. doi:10.1038/leu.2009.38.

Hodge, Victoria J., and Jim Austin. 2004. "A Survey of Outlier Detection Methodologies." *Artificial Intelligence Review* 22 (2). Kluwer Academic Publishers: 85–126. doi:10.1007/s10462-004-4304-y.

Hodis, Eran, Ian R. Watson, Gregory V. Kryukov, Stefan T. Arold, Marcin Imielinski, Jean-Philippe Theurillat, Elizabeth Nickerson, et al. 2012. "A Landscape of Driver Mutations in Melanoma." *Cell* 150 (2): 251–63. doi:10.1016/j.cell.2012.06.024.

Hold, Georgina L., Charles S. Rabkin, Wong–Ho Chow, Malcolm G. Smith, Marilie D. Gammon, Harvey A. Risch, Thomas L. Vaughan, et al. 2007. "A Functional Polymorphism of Toll-Like Receptor 4 Gene Increases Risk of Gastric Carcinoma and Its Precursors." *Gastroenterology* 132 (3): 905–12. doi:10.1053/j.gastro.2006.12.026.

Hopfield, J J. 1982. "Neural Networks and Physical Systems with Emergent Collective Computational Abilities." *Proceedings of the National Academy of Sciences of the United States of America* 79 (8). National Academy of Sciences: 2554–58. doi:10.1073/PNAS.79.8.2554.

Hornbeck, P. V., J. M. Kornhauser, S. Tkachev, B. Zhang, E. Skrzypek, B. Murray, V. Latham, and M. Sullivan. 2012. "PhosphoSitePlus: A Comprehensive Resource for Investigating the Structure and Function of Experimentally Determined Post-Translational Modifications in Man and Mouse." *Nucleic Acids Research* 40 (D1): D261–70. doi:10.1093/nar/gkr1122.

Horowitz, J M, S H Park, E Bogenmann, J C Cheng, D W Yandell, F J Kaye, J D Minna, T P Dryja, and R A Weinberg. 1990. "Frequent Inactivation of the Retinoblastoma Anti-Oncogene Is Restricted to a Subset of Human Tumor Cells." *Proceedings of the National Academy of Sciences of the United States of America* 87 (7). National Academy of Sciences: 2775–79. http://www.ncbi.nlm.nih.gov/pubmed/2181449.

Hsu, Sheng-Da, Yu-Ting Tseng, Sirjana Shrestha, Yu-Ling Lin, Anas Khaleel,

Chih-Hung Chou, Chao-Fang Chu, et al. 2014. “MiRTarBase Update 2014: An Information Resource for Experimentally Validated MiRNA-Target Interactions.” *Nucleic Acids Research* 42 (D1): D78–85. doi:10.1093/nar/gkt1266.

Hung, Jung-Jyh, Chung-Tsen Hsueh, Kuan-Hua Chen, Wen-Hu Hsu, and Yu-Chung Wu. 2012. “Clinical Significance of E2F1 Protein Expression in Non-Small Cell Lung Cancer.” *Experimental Hematology & Oncology* 1 (1): 18. doi:10.1186/2162-3619-1-18.

Hunter, T, and B M Sefton. 1980. “Transforming Gene Product of Rous Sarcoma Virus Phosphorylates Tyrosine.” *Proceedings of the National Academy of Sciences of the United States of America* 77 (3). National Academy of Sciences: 1311–15. http://www.ncbi.nlm.nih.gov/pubmed/6246487.

Husseinzadeh, Nader, and Sara Madison Davenport. 2014. “Role of Toll-like Receptors in Cervical, Endometrial and Ovarian Cancers: A Review.” *Gynecologic Oncology* 135 (2): 359–63. doi:10.1016/j.ygyno.2014.08.013.

Hvid-Jensen, Frederik, Lars Pedersen, Asbjørn Mohr Drewes, Henrik Toft Sørensen, and Peter Funch-Jensen. 2011. “Incidence of Adenocarcinoma among Patients with Barrett’s Esophagus.” *New England Journal of Medicine* 365 (15). Massachusetts Medical Society: 1375–83. doi:10.1056/NEJMoa1103042.

International Cancer Genome Consortium, T J Hudson, W Anderson, A Artez, A D Barker, C Bell, R R Bernabe, et al. 2010. “International Network of Cancer Genome Projects.” *Nature* 464 (7291): 993–98. doi:10.1038/nature08987.

International Human Genome Sequencing Consortium. 2004. “Finishing the Euchromatic Sequence of the Human Genome.” *Nature* 431 (7011): 931–45. doi:10.1038/nature03001.

Iranzo, Jaime, Iñigo Martincorena, and Eugene V Koonin. 2017. “The Cancer-Mutation Network and the Number and Specificity of Driver Mutations.” *BioRxiv*. http://biorxiv.org/content/early/2017/12/19/237016.abstract.

Ishida, S., E. Huang, H. Zuzan, R. Spang, G. Leone, M. West, and J. R. Nevins. 2001. “Role for E2F in Control of Both DNA Replication and Mitotic Functions as Revealed from DNA Microarray Analysis.” *Molecular and Cellular Biology* 21 (14): 4684–99. doi:10.1128/MCB.21.14.4684-

4699.2001.

Issaenko, O A, P B Bitterman, V A Polunovsky, and P S Dahlberg. 2012. “Cap-Dependent {mRNA} Translation and the Ubiquitin-Proteasome System Cooperate to Promote {ERBB2-Dependent} Esophageal Cancer Phenotype.” *Cancer Gene Ther.* 19 (9): 609–18.

Jain, A., and D. Zongker. 1997. “Feature Selection: Evaluation, Application, and Small Sample Performance.” *IEEE Transactions on Pattern Analysis and Machine Intelligence* 19 (2). IEEE Computer Society: 153–58. doi:10.1109/34.574797.

James, Gareth (Gareth Michael). 2013. *An Introduction to Statistical Learning : With Applications in R*. Springer, New York, NY.

Jefford, Charles Edward, and Irmgard Irminger-Finger. 2006. “Mechanisms of Chromosome Instability in Cancers.” *Critical Reviews in Oncology/Hematology* 59 (1): 1–14. doi:10.1016/j.critrevonc.2006.02.005.

Jeong, H., S. P. Mason, A.-L. Barabási, and Z. N. Oltvai. 2001. “Lethality and Centrality in Protein Networks.” *Nature* 411 (6833). Nature Publishing Group: 41–42. doi:10.1038/35075138.

Jewer, Michael, Scott D Findlay, and Lynne-Marie Postovit. 2012. “Post-Transcriptional Regulation in Cancer Progression : Microenvironmental Control of Alternative Splicing and Translation.” *Journal of Cell Communication and Signaling* 6 (4). Springer: 233–48. doi:10.1007/s12079-012-0179-x.

Jiang, Yuqiu, and Mu Wang. 2010. “Personalized Medicine in Oncology: Tailoring the Right Drug to the Right Patient.” *Biomarkers in Medicine* 4 (4): 523–33. doi:10.2217/bmm.10.66.

Johnson, David G. 2000. “The Paradox OfE2F1: Oncogene and Tumor Suppressor Gene.” *Molecular Carcinogenesis* 27 (3). Wiley-Blackwell: 151–57. doi:10.1002/(SICI)1098-2744(200003)27:3<151::AID-MC1>3.0.CO;2-C.

Jones, Chris, Maria Rodriguez-Pinilla, Maryou Lambros, Dorine Bax, Boo Messahel, Gordan M Vujanic, Jorge S Reis-Filho, and Kathy Pritchard-Jones. 2007. “C-KIT Overexpression, without Gene Amplification and Mutation, in Paediatric Renal Tumours.” *Journal of Clinical Pathology* 60 (11). BMJ Publishing Group: 1226–31. doi:10.1136/jcp.2007.046441.

Jonsson, P F, and P A Bates. 2006. “Global Topological Features of Cancer Proteins in the Human Interactome.” *Bioinformatics* 22 (18): 2291–97.

doi:10.1093/bioinformatics/btl390.

Jonsson, Pall F, Tamara Cavanna, Daniel Zicha, and Paul A Bates. 2006. “Cluster Analysis of Networks Generated through Homology: Automatic Identification of Important Protein Communities Involved in Cancer Metastasis.” *BMC Bioinformatics* 7 (1). BioMed Central: 2. doi:10.1186/1471-2105-7-2.

Jusakul, A, I Cutcutache, C H Yong, J Q Lim, M N Huang, N Padmanabhan, V Nellore, et al. 2017. “Whole-Genome and Epigenomic Landscapes of Etiologically Distinct Subtypes of Cholangiocarcinoma.” *Cancer Discov* 7 (10): 1116–35. doi:10.1158/2159-8290.CD-17-0368.

Kadakia, Shailesh C., Henry Renom De la Baume, and Richard T. Shaffer. 1996. “Effects of Transdermal Nicotine on Lower Esophageal Sphincter and Esophageal Motility.” *Digestive Diseases and Sciences* 41 (11). Kluwer Academic Publishers-Plenum Publishers: 2130–34. doi:10.1007/BF02071391.

Kaiser, J. 2008. “GENOMICS: Billion-Dollar Cancer Mapping Project Steps Forward.” *Science* 321 (5885): 26a–27a. doi:10.1126/science.321.5885.26a.

Kandoth, C, M D McLellan, F Vandin, K Ye, B Niu, C Lu, M Xie, et al. 2013. “Mutational Landscape and Significance across 12 Major Cancer Types.” *Nature* 502 (7471): 333–39. doi:10.1038/nature12634.

Kapil, Shweta, Bal Krishan Sharma, Mallikarjun Patil, Sawsan Elattar, Jinling Yuan, Steven X. Hou, Ravindra Kolhe, et al. 2017. “The Cell Polarity Protein Scrib Functions as a Tumor Suppressor in Liver Cancer.” *Oncotarget* 8 (16). Impact Journals: 26515–31. doi:10.18632/oncotarget.15713.

Kelly, Timothy J., Carles Lerin, Wilhelm Haas, Steven P. Gygi, and Pere Puigserver. 2009. “GCN5-Mediated Transcriptional Control of the Metabolic Coactivator PGC-1β through Lysine Acetylation.” *Journal of Biological Chemistry* 284 (30): 19945–52. doi:10.1074/jbc.M109.015164.

Kendall, Bradley J., Joel H. Rubenstein, Michael B. Cook, Thomas L. Vaughan, Lesley A. Anderson, Liam J. Murray, Nicholas J. Shaheen, et al. 2016. “Inverse Association Between Gluteofemoral Obesity and Risk of Barrett’s Esophagus in a Pooled Analysis.” *Clinical Gastroenterology and Hepatology* 14 (10): 1412–1419.e3. doi:10.1016/j.cgh.2016.05.032.

Keshava Prasad, T S, Renu Goel, Kumaran Kandasamy, Shivakumar Keerthikumar, Sameer Kumar, Suresh Mathivanan, Deepthi Telikicherla, et al. 2009. “Human Protein Reference Database--2009 Update.” *Nucleic Acids Research*. doi:10.1093/nar/gkn892.

Khan, Javed, Jun S. Wei, Markus Ringnér, Lao H. Saal, Marc Ladanyi, Frank Westermann, Frank Berthold, et al. 2001. “Classification and Diagnostic Prediction of Cancers Using Gene Expression Profiling and Artificial Neural Networks.” *Nature Medicine* 7 (6): 673–79. doi:10.1038/89044.

Knudson Jr., A G. 1971. “Mutation and Cancer: Statistical Study of Retinoblastoma.” *Proc Natl Acad Sci U S A* 68 (4): 820–23. https://www.ncbi.nlm.nih.gov/pubmed/5279523.

Koike, K, T Tsutsumi, H Miyoshi, S Shinzawa, Y Shintani, H Fujie, H Yotsuyanagi, and K Moriya. 2008. “Molecular Basis for the Synergy between Alcohol and Hepatitis C Virus in Hepatocarcinogenesis.” *J Gastroenterol Hepatol* 23 Suppl 1: S87-91. doi:10.1111/j.1440-1746.2007.05292.x.

Kong, C. Y., K. J. Nattinger, T. J. Hayeck, Z. B. Omer, Y. C. Wang, S. J. Spechler, P. M. McMahon, G. S. Gazelle, and C. Hur. 2011. “The Impact of Obesity on the Rise in Esophageal Adenocarcinoma Incidence: Estimates from a Disease Simulation Model.” *Cancer Epidemiology Biomarkers & Prevention* 20 (11): 2450–56. doi:10.1158/1055-9965.EPI-11-0547.

Koonin, E V. 1993. “A Common Set of Conserved Motifs in a Vast Variety of Putative Nucleic Acid-Dependent ATPases Including MCM Proteins Involved in the Initiation of Eukaryotic DNA Replication.” *Nucleic Acids Research* 21 (11): 2541–47. http://www.ncbi.nlm.nih.gov/pubmed/8332451.

Kopp, H G, and R D Hofheinz. 2016. “Targeted Treatment of Esophagogastric Cancer.” *Oncology Research and Treatment* 39 (12): 788–94.

Kotsiantis, S B. 2007. “Supervised Machine Learning: A Review of Classification Techniques.” *Informatica*. Vol. 31. https://datajobs.com/data-science-repo/Supervised-Learning-[SB-Kotsiantis].pdf.

Krell, Jonathan, Justin Stebbing, Claudia Carissimi, Aleksandra F Dabrowska, Alexander de Giorgio, Adam E Frampton, Victoria Harding, et al. 2016. “TP53 Regulates MiRNA Association with AGO2 to Remodel the MiRNA-MRNA Interaction Network.” *Genome Research* 26 (3). Cold Spring Harbor Laboratory Press: 331–41. doi:10.1101/gr.191759.115.

Kroep, S, I Lansdorp-Vogelaar, J H Rubenstein, V E P P Lemmens, E B van Heijningen, N Aragonés, M van Ballegooijen, and J M Inadomi. 2014. “Comparing Trends in Esophageal Adenocarcinoma Incidence and Lifestyle Factors Between the United States, Spain and The Netherlands.” *The American Journal of Gastroenterology* 109 (3): 336–43. doi:10.1038/ajg.2013.420.

Krutzik, Peter O, and Garry P Nolan. 2006. “Fluorescent Cell Barcoding in Flow Cytometry Allows High-Throughput Drug Screening and Signaling Profiling.” *Nature Methods* 3 (5): 361–68. doi:10.1038/nmeth872.

Kumar, P, S Henikoff, and P C Ng. 2009. “Predicting the Effects of Coding Non-Synonymous Variants on Protein Function Using the SIFT Algorithm.” *Nat Protoc* 4 (7): 1073–81. doi:10.1038/nprot.2009.86.

Kuncheva, Ludmila I. 2004. *Combining Pattern Classifiers Methods and Algorithms*. John Wiley & Sons, Inc., Hoboken, New Jersey. www.copyright.com.

Kurt, Hulyam, Cansu Ozbayer, Aysegul Bayramoglu, Hasan Veysi Gunes, İrfan Degirmenci, Kevser Setenay Oner, and Muzaffer Metintas. 2016. “Determination of the Relationship Between Rs4986790 and Rs4986791 Variants of TLR4 Gene and Lung Cancer.” *Inflammation* 39 (1): 166–71. doi:10.1007/s10753-015-0235-9.

Kwok, Hang Fai, Shu-Dong Zhang, Cian M McCrudden, Hiu-Fung Yuen, Kam-Po Ting, Qing Wen, Ui-Soon Khoo, and Kelvin Yuen-Kwong Chan. 2015. “Prognostic Significance of Minichromosome Maintenance Proteins in Breast Cancer.” *American Journal of Cancer Research* 5 (1): 52–71. http://www.ncbi.nlm.nih.gov/pubmed/25628920.

Lagergren, Jesper. 2011. “Influence of Obesity on the Risk of Esophageal Disorders.” *Nature Reviews Gastroenterology & Hepatology* 8 (6): 340–47. doi:10.1038/nrgastro.2011.73.

Lagergren, Jesper, Reinhold Bergström, Anders Lindgren, and Olof Nyrén. 1999. “Symptomatic Gastroesophageal Reflux as a Risk Factor for Esophageal Adenocarcinoma.” *New England Journal of Medicine* 340 (11): 825–31. doi:10.1056/NEJM199903183401101.

Lander, E S, and R A Weinberg. 2000. “Genomics: Journey to the Center of Biology.” *Science (New York, N.Y.)* 287 (5459): 1777–82. http://www.ncbi.nlm.nih.gov/pubmed/10755930.

Lane, D P, and L V Crawford. 1979. “T Antigen Is Bound to a Host Protein in SV40-Transformed Cells.” *Nature* 278 (5701): 261–63. https://www.ncbi.nlm.nih.gov/pubmed/218111.

Lawrence, M S, P Stojanov, P Polak, G V Kryukov, K Cibulskis, A Sivachenko, S L Carter, et al. 2013. “Mutational Heterogeneity in Cancer and the Search for New Cancer-Associated Genes.” *Nature* 499 (7457): 214–18. doi:10.1038/nature12213.

Lawrence, Michael S., Petar Stojanov, Craig H. Mermel, James T. Robinson, Levi A. Garraway, Todd R. Golub, Matthew Meyerson, Stacey B. Gabriel, Eric S. Lander, and Gad Getz. 2014. “Discovery and Saturation Analysis of Cancer Genes across 21 Tumour Types.” *Nature* 505 (7484): 495–501. doi:10.1038/nature12912.

Lawrence, Michael S, Petar Stojanov, Paz Polak, Gregory V Kryukov, Kristian Cibulskis, Andrey Sivachenko, Scott L Carter, et al. 2013. “Mutational Heterogeneity in Cancer and the Search for New Cancer-Associated Genes.” doi:10.1038/nature12213.

Lee, Chang-Yong. 2006. “Correlations among Centrality Measures in Complex Networks.” https://arxiv.org/pdf/physics/0605220.pdf.

Leiserson, M D, F Vandin, H T Wu, J R Dobson, J V Eldridge, J L Thomas, A Papoutsaki, et al. 2015. “Pan-Cancer Network Analysis Identifies Combinations of Rare Somatic Mutations across Pathways and Protein Complexes.” *Nat Genet* 47 (2): 106–14. doi:10.1038/ng.3168.

Lengauer, Christoph, Kenneth W. Kinzler, and Bert Vogelstein. 1998. “Genetic Instabilities in Human Cancers.” *Nature* 396 (6712): 643–49. doi:10.1038/25292.

Levine, A J. 1997. “P53, the Cellular Gatekeeper for Growth and Division.” *Cell* 88 (3): 323–31. https://www.ncbi.nlm.nih.gov/pubmed/9039259.

Li, Shutao, James T. Kwok, Hailong Zhu, and Yaonan Wang. 2003. “Texture Classification Using the Support Vector Machines.” *Pattern Recognition* 36 (12). Pergamon: 2883–93. doi:10.1016/S0031-3203(03)00219-X.

Li, Siming, Christopher M Armstrong, Nicolas Bertin, Hui Ge, Stuart Milstein, Mike Boxem, Pierre-Olivier Vidalain, et al. 2004. “A Map of the Interactome Network of the Metazoan C. Elegans.” *Science (New York, N.Y.)* 303 (5657). American Association for the Advancement of Science: 540–43. doi:10.1126/science.1091403.

Li, Xiaoli, and Bing Liu. 2003. “Learning to Classify Texts Using Positive and Unlabeled Data.” *Proceedings of the 18th International Joint Conference on Artificial Intelligence*. Morgan Kaufmann Publishers Inc. https://dl.acm.org/citation.cfm?id=1630746.

Li, Xin, and Dave Thirumalai. 2016. “Interplay of Driver, Mini-Driver, and Deleterious Passenger Mutations on Cancer Progression.” *BioRxiv*, October. Cold Spring Harbor Laboratory, 084392. doi:10.1101/084392.

Li, Yifeng, Fang-Xiang Wu, and Alioune Ngom. 2016. “A Review on Machine Learning Principles for Multi-View Biological Data Integration.” *Briefings in Bioinformatics* 19 (2). Oxford University Press: bbw113. doi:10.1093/bib/bbw113.

Liang, Yu-Xiang, Jian-Ming Lu, Ru-Jun Mo, Hui-Chan He, Jian Xie, Fu-Neng Jiang, Zhuo-Yuan Lin, et al. 2016. “E2F1 Promotes Tumor Cell Invasion and Migration through Regulating CD147 in Prostate Cancer.” *International Journal of Oncology* 48 (4). Spandidos Publications: 1650–58. doi:10.3892/ijo.2016.3364.

Lieberman-Aiden, E, N L van Berkum, L Williams, M Imakaev, T Ragoczy, A Telling, I Amit, et al. 2009. “Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome.” *Science* 326 (5950): 289–93.

Lièvre, Astrid, Jean-Baptiste Bachet, Delphine Le Corre, Valérie Boige, Bruno Landi, Jean-François Emile, Jean-François Côté, et al. 2006. “*KRAS* Mutation Status Is Predictive of Response to Cetuximab Therapy in Colorectal Cancer.” *Cancer Research* 66 (8): 3992–95. doi:10.1158/0008-5472.CAN-06-0191.

Lindsay, Helen, Alexa Burger, Berthin Biyong, Anastasia Felker, Christopher Hess, Jonas Zaugg, Elena Chiavacci, et al. 2016. “{CrispRVariants} Charts the Mutation Spectrum of Genome Engineering Experiments.” *Nat. Biotechnol.* 34 (7): 701–2.

Little, A. S., K. Balmanno, M. J. Sale, S. Newman, J. R. Dry, M. Hampson, P. A. W. Edwards, P. D. Smith, and S. J. Cook. 2011. “Amplification of the Driving Oncogene, KRAS or BRAF, Underpins Acquired Resistance to MEK1/2 Inhibitors in Colorectal Cancer Cells.” *Science Signaling* 4 (166): ra17-ra17. doi:10.1126/scisignal.2001752.

Liu, Fei Tony, Kai Ming Ting, and Zhi-hua Zhou. 2008. “Isolation Forest.” In

Proceedings of the 2008 eighth IEEE International Conference on Data Mining. IEEE Computer Society, 413–22. doi=10.1.1.678.3903.

Liu, X, C Wu, C Li, and E Boerwinkle. 2016. “dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs.” *Hum. Mutat.* 37 (3): 235–41.

Liu, Yi-Zhen, Bo-Shi Wang, Yan-Yi Jiang, Jian Cao, Jia-Jie Hao, Yu Zhang, Xin Xu, Yan Cai, and Ming-Rong Wang. 2017. “MCMs Expression in Lung Cancer: Implication of Prognostic Significance.” *Journal of Cancer* 8 (18). Ivyspring International Publisher: 3641–47. doi:10.7150/jca.20777.

Loeb, L A, and C C Harris. 2008. “Advances in Chemical Carcinogenesis: A Historical Review and Prospective.” *Cancer Res* 68 (17): 6863–72. doi:10.1158/0008-5472.CAN-08-2852.

Lonsdale, John, Jeffrey Thomas, Mike Salvatore, Rebecca Phillips, Edmund Lo, Saboor Shad, Richard Hasz, et al. 2013. “The Genotype-Tissue Expression (GTEx) Project.” *Nature Genetics*. doi:10.1038/ng.2653.

Loo, P Van, S H Nordgard, O C Lingjaerde, H G Russnes, I H Rye, W Sun, V J Weigman, et al. 2010. “Allele-Specific Copy Number Analysis of Tumors.” *Proc. Natl. Acad. Sci. U. S. A.* 107 (39): 16910–15.

Lord, Christopher J, and Alan Ashworth. 2017. “PARP Inhibitors: Synthetic Lethality in the Clinic.” *Science (New York, N.Y.)* 355 (6330). American Association for the Advancement of Science: 1152–58. doi:10.1126/science.aam7344.

Lu, Meixia, Zhensheng Liu, Hongping Yu, Li-E Wang, Guojun Li, Erich M. Sturgis, David G. Johnson, and Qingyi Wei. 2012. “Combined Effects of *E2F1* and *E2F2* Polymorphisms on Risk and Early Onset of Squamous Cell Carcinoma of the Head and Neck.” *Molecular Carcinogenesis* 51 (S1): E132–41. doi:10.1002/mc.21882.

Lukas, J, B O Petersen, K Holm, J Bartek, and K Helin. 1996. “Deregulated Expression of {E2F} Family Members Induces S-Phase Entry and Overcomes {p16INK4A-Mediated} Growth Suppression.” *Mol. Cell. Biol.* 16 (3): 1047–57.

Lydon, Nicholas B, and Brian J Druker. 2004. “Lessons Learned from the Development of Imatinib.” *Leukemia Research* 28 (May): 29–38. doi:10.1016/j.leukres.2003.10.002.

Makino, Takashi, Aoife McLysaght, and Masakado Kawata. 2013. “Genome-

Wide Deserts for Copy Number Variation in Vertebrates." *Nature Communications*. doi:10.1038/ncomms3283.

Maley, Carlo C., Patricia C. Galipeau, Xiaohong Li, Carissa A. Sanchez, Thomas G. Paulson, and Brian J. Reid. 2004. "Selectively Advantageous Mutations and Hitchhikers in Neoplasms." *Cancer Research* 64 (10): 3414–27. doi:10.1158/0008-5472.CAN-03-3249.

Manurung, Jonson, Herman Mawengkang, and Elviawaty Zamzami. 2017. "Optimizing Support Vector Machine Parameters with Genetic Algorithm for Credit Risk Assessment." *Journal of Physics: Conference Series* 930 (1). IOP Publishing: 012026. doi:10.1088/1742-6596/930/1/012026.

Marchler-Bauer, A, C Zheng, F Chitsaz, M K Derbyshire, L Y Geer, R C Geer, N R Gonzales, et al. 2013. "{CDD}: Conserved Domains and Protein Three-Dimensional Structure." *Nucleic Acids Res.* 41 (Database issue): D348--52.

Marchler-Bauer, Aron, Shennan Lu, John B. Anderson, Farideh Chitsaz, Myra K. Derbyshire, Carol DeWeese-Scott, Jessica H. Fong, et al. 2011. "CDD: A Conserved Domain Database for the Functional Annotation of Proteins." *Nucleic Acids Research*. doi:10.1093/nar/gkq1189.

Martin, G S. 1970. "Rous Sarcoma Virus: A Function Required for the Maintenance of the Transformed State." *Nature* 227 (5262): 1021–23. http://www.ncbi.nlm.nih.gov/pubmed/4317808.

Martincorena, I., A. Roshan, M. Gerstung, P. Ellis, P. Van Loo, S. McLaren, D. C. Wedge, et al. 2015. "High Burden and Pervasive Positive Selection of Somatic Mutations in Normal Human Skin." *Science* 348 (6237): 880–86. doi:10.1126/science.aaa6806.

Martincorena, I, K M Raine, M Gerstung, K J Dawson, K Haase, P Van Loo, H Davies, M R Stratton, and P J Campbell. 2017. "Universal Patterns of Selection in Cancer and Somatic Tissues." *Cell* 171 (5): 1029–1041 e21. doi:10.1016/j.cell.2017.09.042.

Mathur, Sunil, and Joseph Sutton. 2017. "Personalized Medicine Could Transform Healthcare." *Biomedical Reports* 7 (1). Spandidos Publications: 3–5. doi:10.3892/br.2017.922.

McFarland, C D, K S Korolev, G V Kryukov, S R Sunyaev, and L A Mirny. 2013. "Impact of Deleterious Passenger Mutations on Cancer Progression." *Proc Natl Acad Sci U S A* 110 (8): 2910–15. doi:10.1073/pnas.1213968110.

McKenna, A, M Hanna, E Banks, A Sivachenko, K Cibulskis, A Kernytsky, K

Garimella, et al. 2010. “The Genome Analysis Toolkit: A {MapReduce} Framework for Analyzing next-Generation {DNA} Sequencing Data.” *Genome Res.* 20 (9): 1297–1303.

Meads, M B, R A Gatenby, and W S Dalton. 2009. “Environment-Mediated Drug Resistance: A Major Contributor to Minimal Residual Disease.” *Nat Rev Cancer* 9 (9): 665–74. doi:10.1038/nrc2714.

Mele, M, P G Ferreira, F Reverter, D S DeLuca, J Monlong, M Sammeth, T R Young, et al. 2015. “Human Genomics. The Human Transcriptome across Tissues and Individuals.” *Science* 348 (6235): 660–65.

Meyerson, Matthew, Stacey Gabriel, and Gad Getz. 2010. “Advances in Understanding Cancer Genomes through Second-Generation Sequencing.” *Nature Reviews Genetics* 11 (10). Nature Publishing Group: 685–96. doi:10.1038/nrg2841.

Milacic, Marija, Robin Haw, Karen Rothfels, Guanming Wu, David Croft, Henning Hermjakob, Peter D’Eustachio, and Lincoln Stein. 2012. “Annotating Cancer Variants and Anti-Cancer Therapeutics in Reactome.” *Cancers*. doi:10.3390/cancers4041180.

Miller, D G. 1980. “On the Nature of Susceptibility to Cancer. The Presidential Address.” *Cancer* 46 (6): 1307–18. https://www.ncbi.nlm.nih.gov/pubmed/7417931.

Miller, Elizabeth C, and James A Miller. 1947. “The Presence and Significance of Bound Aminoazo Dyes in the Livers of Rats Fed P-Dimethylaminoazobenzene*.” *Cancer Res.* 7: 468–80. http://cancerres.aacrjournals.org/content/canres/7/7/468.full.pdf.

Misale, Sandra, Rona Yaeger, Sebastijan Hobor, Elisa Scala, Manickam Janakiraman, David Liska, Emanuele Valtorta, et al. 2012. “Emergence of KRAS Mutations and Acquired Resistance to Anti-EGFR Therapy in Colorectal Cancer.” *Nature* 486 (7404): 532–36. doi:10.1038/nature11156.

Mok, Tony S.K., Yi-Long Wu, Chong-Jen Yu, Caicun Zhou, Yuh-Min Chen, Li Zhang, Jorge Ignacio, et al. 2009. “Randomized, Placebo-Controlled, Phase II Study of Sequential Erlotinib and Chemotherapy As First-Line Treatment for Advanced Non–Small-Cell Lung Cancer.” *Journal of Clinical Oncology* 27 (30): 5080–87. doi:10.1200/JCO.2008.21.5541.

Molenaar, Jan J., Jan Koster, Danny A. Zwijnenburg, Peter van Sluis, Linda J. Valentijn, Ida van der Ploeg, Mohamed Hamdi, et al. 2012. “Sequencing of

Neuroblastoma Identifies Chromothripsis and Defects in Neuritogenesis Genes." *Nature* 483 (7391): 589–93. doi:10.1038/nature10910.

Montenegro, María F, María del Mar Collado-González, María Piedad Fernández-Pérez, Manel B Hammouda, Lana Tolordava, Mariam Gamkrelidze, and José Neptuno Rodríguez-López. 2014. "Promoting E2F1-Mediated Apoptosis in Oestrogen Receptor-α-Negative Breast Cancer Cells." *BMC Cancer* 14 (1): 539. doi:10.1186/1471-2407-14-539.

Mordelet, Fantine, and Jean-Philippe Vert. 2010. "A Bagging SVM to Learn from Positive and Unlabeled Examples." https://hal.archives-ouvertes.fr/hal-00523336.

Moulder, S. 2010. "Intrinsic Resistance to Chemotherapy in Breast Cancer." *Womens Health (Lond)* 6 (6): 821–30. doi:10.2217/whe.10.60.

Moya, M. M., M. W. Koch, and L. D. Hostetler. 1993. "One-Class Classifier Networks for Target Recognition Applications." In *In Proceedings World Congress on Neural Networks*, 797–801. http://adsabs.harvard.edu/abs/1993STIN...9324043M.

Mwangi, Benson, Tian Siva Tian, and Jair C Soares. 2014. "A Review of Feature Reduction Techniques in Neuroimaging." *Neuroinformatics* 12 (2). NIH Public Access: 229–44. doi:10.1007/s12021-013-9204-3.

Nam, Su Youn, Il Ju Choi, Kum Hei Ryu, Bum Joon Park, Hyun Bum Kim, and Byung–Ho Nam. 2010. "Abdominal Visceral Adipose Tissue Volume Is Associated With Increased Risk of Erosive Esophagitis in Men and Women." *Gastroenterology* 139 (6): 1902–1911.e2. doi:10.1053/j.gastro.2010.08.019.

Nath, S, A Chowdhury, S Dey, A Roychoudhury, A Ganguly, D Bhattacharyya, and S Roychoudhury. 2015. "Deregulation of {Rb-E2F1} Axis Causes Chromosomal Instability by Engaging the Transactivation Function of Cdc20-Anaphase-Promoting Complex/Cyclosome." *Mol. Cell. Biol.* 35 (2): 356–69.

Cancer Genome Atlas Research Network, T J Ley, C Miller, L Ding, B J Raphael, A J Mungall, A Robertson, et al. 2013. "Genomic and Epigenomic Landscapes of Adult de Novo Acute Myeloid Leukemia." *N Engl J Med* 368 (22): 2059–74. doi:10.1056/NEJMoa1301689.

Nik-Zainal, Serena, Ludmil B. Alexandrov, David C. Wedge, Peter Van Loo, Christopher D. Greenman, Keiran Raine, David Jones, et al. 2012.

"Mutational Processes Molding the Genomes of 21 Breast Cancers." *Cell* 149 (5): 979–93. doi:10.1016/j.cell.2012.04.024.

Nishikura, Kazuko. 2016. "A-to-I Editing of Coding and Non-Coding RNAs by ADARs." *Nature Reviews Molecular Cell Biology* 17 (2). Nature Publishing Group: 83–96. doi:10.1038/nrm.2015.4.

Noble, William S. 2006. "What Is a Support Vector Machine?" *Nature Biotechnology*. Vol. 24.

Nones, Katia, Nicola Waddell, Nicci Wayte, Ann-Marie Patch, Peter Bailey, Felicity Newell, Oliver Holmes, et al. 2014. "Genomic Catastrophes Frequently Arise in Esophageal Adenocarcinoma and Drive Tumorigenesis." *Nature Communications* 5 (1). Nature Publishing Group: 5224. doi:10.1038/ncomms6224.

Nordenstedt, Helena, and Hashem El-Serag. 2011. "The Influence of Age, Sex, and Race on the Incidence of Esophageal Cancer in the United States (1992–2006)." *Scandinavian Journal of Gastroenterology* 46 (5): 597–602. doi:10.3109/00365521.2011.551890.

Nougarède, R, F Della Seta, P Zarzov, and E Schwob. 2000. "Hierarchy of S-Phase-Promoting Factors: Yeast Dbf4-Cdc7 Kinase Requires Prior S-Phase Cyclin-Dependent Kinase Activation." *Molecular and Cellular Biology* 20 (11): 3795–3806. http://www.ncbi.nlm.nih.gov/pubmed/10805723.

Nowell, P C. 1976. "The Clonal Evolution of Tumor Cell Populations." *Science* 194 (4260): 23–28. https://www.ncbi.nlm.nih.gov/pubmed/959840.

Nuttall, Srewart D., Brendon J. Hanson, Masataka Mori, and Nicholas J. Hoogenraad. 1997. "HTom34: A Novel Translocase for the Import of Proteins into Human Mitochondria." *DNA and Cell Biology* 16 (9): 1067–74. doi:10.1089/dna.1997.16.1067.

Ogawa, H., Kei-Ichiro Ishiguro, Stefan Gaubatz, David M Livingston, and Yoshihiro Nakatani. 2002. "A Complex with Chromatin Modifiers That Occupies E2F- and Myc-Responsive Genes in G0 Cells." *Science* 296 (5570): 1132–36. doi:10.1126/science.1069861.

Ohtani, K, R Iwanaga, M Nakamura, M Ikeda, N Yabuta, H Tsuruga, and H Nojima. 1999. "Cell Growth-Regulated Expression of Mammalian {MCM5} and {MCM6} Genes Mediated by the Transcription Factor {E2F}." *Oncogene* 18 (14): 2299–2309.

Olivier, M, M Hollstein, and P Hainaut. 2010. "TP53 Mutations in Human

Cancers: Origins, Consequences, and Clinical Use." *Cold Spring Harb Perspect Biol* 2 (1): a001008. doi:10.1101/cshperspect.a001008.

Olliver, J. R., Laura J Hardie, Yunyun Gong, Simon Dexter, Douglas Chalmers, Keith M Harris, and Christopher P Wild. 2005. "Risk Factors, DNA Damage, and Disease Progression in Barrett's Esophagus." *Cancer Epidemiology Biomarkers & Prevention* 14 (3): 620–25. doi:10.1158/1055-9965.EPI-04-0509.

Olsen, C. M., N. Pandeya, A. C. Green, P. M. Webb, D. C. Whiteman, and Australian Cancer Study. 2011. "Population Attributable Fractions of Adenocarcinoma of the Esophagus and Gastroesophageal Junction." *American Journal of Epidemiology* 174 (5): 582–90. doi:10.1093/aje/kwr117.

Orchard, Sandra, Mais Ammari, Bruno Aranda, Lionel Breuza, Leonardo Briganti, Fiona Broackes-Carter, Nancy H. Campbell, et al. 2014. "The MIntAct Project—IntAct as a Common Curation Platform for 11 Molecular Interaction Databases." *Nucleic Acids Research* 42 (D1): D358–63. doi:10.1093/nar/gkt1115.

Orpinell, Meritxell, Marjorie Fournier, Anne Riss, Zita Nagy, Arnaud R Krebs, Mattia Frontini, and Làszlò Tora. 2010. "The ATAC Acetyl Transferase Complex Controls Mitotic Progression by Targeting Non-Histone Substrates." *The EMBO Journal* 29 (14): 2381–94. doi:10.1038/emboj.2010.125.

Pai, Chen-Chun, Rachel S. Deegan, Lakxmi Subramanian, Csenge Gal, Sovan Sarkar, Elizabeth J. Blaikley, Carol Walker, et al. 2014. "A Histone H3K36 Chromatin Switch Coordinates DNA Double-Strand Break Repair Pathway Choice." *Nature Communications* 5 (1): 4091. doi:10.1038/ncomms5091.

Pal, Nikhil R, Kripamoy Aguan, Animesh Sharma, and Shun-ichi Amari. 2007. "Discovering Biomarkers from Gene Expression Data for Predicting Cancer Subgroups Using Neural Networks and Relational Fuzzy Clustering." *BMC Bioinformatics* 8 (1). BioMed Central: 5. doi:10.1186/1471-2105-8-5.

Pandeya, Nirmala, Catherine M. Olsen, and David C. Whiteman. 2013. "Sex Differences in the Proportion of Esophageal Squamous Cell Carcinoma Cases Attributable to Tobacco Smoking and Alcohol Consumption." *Cancer Epidemiology* 37 (5). Elsevier: 579–84. doi:10.1016/J.CANEP.2013.05.011.

Paolinelli, Roberta, Ramiro Mendoza-Maldonado, Anna Cereseto, and Mauro

Giacca. 2009. “Acetylation by GCN5 Regulates CDC6 Phosphorylation in the S Phase of the Cell Cycle.” *Nature Structural & Molecular Biology* 16 (4): 412–20. doi:10.1038/nsmb.1583.

Pardo, Eduard Porta, and Adam Godzik. 2015. “Analysis of Individual Protein Regions Provides Novel Insights on Cancer Pharmacogenomics.” Edited by Christine A. Orengo. *PLoS Computational Biology* 11 (1): e1004024. doi:10.1371/journal.pcbi.1004024.

Peng, Xinxin, Xiaoyan Xu, Yumeng Wang, David H. Hawke, Shuangxing Yu, Leng Han, Zhicheng Zhou, et al. 2018. “A-to-I RNA Editing Contributes to Proteomic Diversity in Cancer.” *Cancer Cell* 33 (5). Cell Press: 817–828.e7. doi:10.1016/J.CCELL.2018.03.026.

Pennathur, Arjun, Michael K Gibson, Blair A Jobe, and James D Luketich. 2013. “Oesophageal Carcinoma.” *The Lancet* 381 (9864). Elsevier: 400–412. doi:10.1016/S0140-6736(12)60643-6.

Peralta-Arrieta, Irlanda, Daniel Hernández-Sotelo, Yaneth Castro-Coronel, Marco Antonio Leyva-Vázquez, and Berenice Illades-Aguiar. 2017. “DNMT3B Modulates the Expression of Cancer-Related Genes and Downregulates the Expression of the Gene VAV3 via Methylation.” *American Journal of Cancer Research* 7 (1). e-Century Publishing Corporation: 77–87. http://www.ncbi.nlm.nih.gov/pubmed/28123849.

Piccart-Gebhart, Martine J., Marion Procter, Brian Leyland-Jones, Aron Goldhirsch, Michael Untch, Ian Smith, Luca Gianni, et al. 2005. “Trastuzumab after Adjuvant Chemotherapy in HER2-Positive Breast Cancer.” *New England Journal of Medicine* 353 (16): 1659–72. doi:10.1056/NEJMoa052306.

Pollard, K S, M J Hubisz, K R Rosenbloom, and A Siepel. 2010. “Detection of Nonneutral Substitution Rates on Mammalian Phylogenies.” *Genome Res* 20 (1): 110–21. doi:10.1101/gr.097857.109.

Pon, J R, and M A Marra. 2015. “Driver and Passenger Mutations in Cancer.” *Annu Rev Pathol* 10: 25–50. doi:10.1146/annurev-pathol-012414-040312.

Porta-Pardo, Eduard, Luz Garcia-Alonso, Thomas Hrabe, Joaquin Dopazo, and Adam Godzik. 2015. “A Pan-Cancer Catalogue of Cancer Driver Protein Interaction Interfaces.” *PLoS Computational Biology* 11 (10). Public Library of Science: e1004518. doi:10.1371/journal.pcbi.1004518.

Porter, Andrew P, Alexandra Papaioannou, and Angeliki Malliri. 2016.

"Deregulation of Rho GTPases in Cancer." *Small GTPases* 7 (3). Taylor & Francis: 123–38. doi:10.1080/21541248.2016.1173767.

Powell, S, K Forslund, D Szklarczyk, K Trachana, A Roth, J Huerta-Cepas, T Gabaldon, et al. 2014. "{eggNOG} v4.0: Nested Orthology Inference across 3686 Organisms." *Nucleic Acids Res.* 42 (Database issue): D231--9.

Pruitt, K D, G R Brown, S M Hiatt, F Thibaud-Nissen, A Astashyn, O Ermolaeva, C M Farrell, et al. 2014. "{RefSeq}: An Update on Mammalian Reference Sequences." *Nucleic Acids Res.* 42 (Database issue): D756--63.

Pudil, P, J Novovieova, and J Kittler. 1994. "Floating Search Methods in Feature Selection." *Pattern Recognition Letters*, no. 15: 1119–25. http://library.utia.cas.cz/separaty/historie/somol-floating search methods in feature selection.pdf.

Qu, Kai, Zhixin Wang, Haining Fan, Juan Li, Jie Liu, Pingping Li, Zheyong Liang, et al. 2017. "MCM7 Promotes Cancer Progression through Cyclin D1-Dependent Signaling and Serves as a Prognostic Marker for Patients with Hepatocellular Carcinoma." *Cell Death & Disease* 8 (2). Nature Publishing Group: e2603–e2603. doi:10.1038/cddis.2016.352.

Ramaswamy, S., P. Tamayo, R. Rifkin, S. Mukherjee, C.-H. Yeang, M. Angelo, C. Ladd, et al. 2001. "Multiclass Cancer Diagnosis Using Tumor Gene Expression Signatures." *Proceedings of the National Academy of Sciences* 98 (26): 15149–54. doi:10.1073/pnas.211566398.

Rambaldi, D, F M Giorgi, F Capuani, A Ciliberto, and F D Ciccarelli. 2008. "Low Duplicability and Network Fragility of Cancer Genes." *Trends Genet* 24 (9): 427–30. doi:10.1016/j.tig.2008.06.003.

Rao, Meghana, Wenqiang Song, Aixiang Jiang, Yu Shyr, Sima Lev, David Greenstein, Dana Brantley-Sieders, and Jin Chen. 2012. "VAMP-Associated Protein B (VAPB) Promotes Breast Tumor Growth by Modulation of Akt Activity." Edited by Waldemar Debinski. *PLoS ONE* 7 (10). Public Library of Science: e46281. doi:10.1371/journal.pone.0046281.

Reddy, E P, R K Reynolds, E Santos, and M Barbacid. 1982. "A Point Mutation Is Responsible for the Acquisition of Transforming Properties by the T24 Human Bladder Carcinoma Oncogene." *Nature* 300 (5888): 149–52. https://www.ncbi.nlm.nih.gov/pubmed/7133135.

Reid, Brian J., Xiaohong Li, Patricia C. Galipeau, and Thomas L. Vaughan. 2010. "Barrett's Oesophagus and Oesophageal Adenocarcinoma: Time for

a New Synthesis." *Nature Reviews Cancer* 10 (2): 87–101. doi:10.1038/nrc2773.

Renan, M J. 1993. "How Many Mutations Are Required for Tumorigenesis? Implications from Human Cancer Data." *Mol Carcinog* 7 (3): 139–46. https://www.ncbi.nlm.nih.gov/pubmed/8489711.

Repana, Dimitra, Joel Nulsen, Lisa Dressler, Michele Bortolomeazzi, Santhilata Kuppili Venkata, Aikaterini Tourna, Anna Yakovleva, Tommaso Palmieri, and Francesca D Ciccarelli. 2018. "The Network of Cancer Genes (NCG): A Comprehensive Catalogue of Known and Candidate Cancer Genes from Cancer Sequencing Screens." *BioRxiv*, December. Cold Spring Harbor Laboratory, 389858. doi:10.1101/389858.

Reva, B, Y Antipin, and C Sander. 2011. "Predicting the Functional Impact of Protein Mutations: Application to Cancer Genomics." *Nucleic Acids Res* 39 (17): e118. doi:10.1093/nar/gkr407.

Ritter, Gunter, and María Teresa Gallegos. 1997. "Outliers in Statistical Pattern Recognition and an Application to Automatic Chromosome Classification." *Pattern Recognition Letters* 18 (6). Elsevier Science Inc.: 525–39. doi:10.1016/S0167-8655(97)00049-4.

Roberts, S.J., W. Penny, and D. Pillot. 1996. "Novelty, Confidence and Errors in Connectionist Systems." In *IEE Colloquium on Intelligent Sensors*, 1996:10–10. IEE. doi:10.1049/ic:19961391.

Rokach, Lior, and Oded Maimon. 2005. "Clustering Methods." In *Data Mining and Knowledge Discovery Handbook*, 321–52. New York: Springer-Verlag. doi:10.1007/0-387-25465-X_15.

Romond, E H, E A Perez, J Bryant, V J Suman, C E Geyer Jr., N E Davidson, E Tan-Chiu, et al. 2005. "Trastuzumab plus Adjuvant Chemotherapy for Operable HER2-Positive Breast Cancer." *N Engl J Med* 353 (16): 1673–84. doi:10.1056/NEJMoa052122.

Ronkainen, Jukka, Pertti Aro, Tom Storskrubb, Sven–Erik Johansson, Tore Lind, Elisabeth Bolling–Sternevald, Michael Vieth, Manfred Stolte, Nicholas J. Talley, and Lars Agréus. 2005. "Prevalence of Barrett's Esophagus in the General Population: An Endoscopic Study." *Gastroenterology* 129 (6): 1825–31. doi:10.1053/j.gastro.2005.08.053.

Ross-Innes, Caryn S, Jennifer Becq, Andrew Warren, R Keira Cheetham, Helen Northen, Maria O'Donovan, Shalini Malhotra, et al. 2015. "Whole-Genome

Sequencing Provides New Insights into the Clonal Architecture of Barrett's Esophagus and Esophageal Adenocarcinoma." *Nat. Genet.* 47. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.: 1038–46.

Roth, Shalom Hillel, Miri Danan-Gotthold, Meirav Ben-Izhak, Gideon Rechavi, Cyrille J. Cohen, Yoram Louzoun, and Erez Y. Levanon. 2018. "Increased RNA Editing May Provide a Source for Autoantigens in Systemic Lupus Erythematosus." *Cell Reports* 23 (1). Cell Press: 50–57. doi:10.1016/J.CELREP.2018.03.036.

Rous, P. 1910. "A Transmissible Avian Neoplasm. (Sarcoma of the Common Fowl.)." *J Exp Med* 12 (5): 696–705. https://www.ncbi.nlm.nih.gov/pubmed/19867354.

Rousseeuw, Peter. 1985. "Multivariate Estimation with High Breakdown Point." In *Mathematical Statistics and Applications*, 283–97. Dordrecht: Springer Netherlands. doi:10.1007/978-94-009-5438-0_20.

Rousseeuw, Peter J. 1987. "Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis." *Journal of Computational and Applied Mathematics* 20 (November). North-Holland: 53–65. doi:10.1016/0377-0427(87)90125-7.

Rowley, J D. 1973. "Letter: A New Consistent Chromosomal Abnormality in Chronic Myelogenous Leukaemia Identified by Quinacrine Fluorescence and Giemsa Staining." *Nature* 243 (5405): 290–93. https://www.ncbi.nlm.nih.gov/pubmed/4126434.

Rubenstein, Joel H., and Nicholas J. Shaheen. 2015. "Epidemiology, Diagnosis, and Management of Esophageal Adenocarcinoma." *Gastroenterology* 149 (2): 302–317.e1. doi:10.1053/j.gastro.2015.04.053.

Ruepp, Andreas, Brigitte Waegele, Martin Lechner, Barbara Brauner, Irmtraud Dunger-Kaltenbach, Gisela Fobo, Goar Frishman, Corinna Montrone, and H. Werner Mewes. 2009. "CORUM: The Comprehensive Resource of Mammalian Protein Complexes-2009." *Nucleic Acids Research*. doi:10.1093/nar/gkp914.

Sabarinathan, Radhakrishnan, Oriol Pich, Inigo Martincorena, Carlota Rubio-Perez, Malene Juul, Jeremiah Wala, Steven Schumacher, et al. 2017. "The Whole-Genome Panorama of Cancer Drivers." *BioRxiv*. doi:10.1101/190330.

Sadoul, Karin, Jin Wang, Boubou Diagouraga, and Saadi Khochbin. 2011. “The Tale of Protein Lysine Acetylation in the Cytoplasm.” *Journal of Biomedicine and Biotechnology* 2011: 1–15. doi:10.1155/2011/970382.

Saldivar, Joshua C., Satoshi Miuma, Jessica Bene, Seyed Ali Hosseini, Hidetaka Shibata, Jin Sun, Linda J. Wheeler, Christopher K. Mathews, and Kay Huebner. 2012. “Initiation of Genome Instability and Preneoplastic Processes through Loss of Fhit Expression.” Edited by Marshall S. Horwitz. *PLoS Genetics* 8 (11). Public Library of Science: e1003077. doi:10.1371/journal.pgen.1003077.

Saletta, Federica, Michaela S Seng, and Loretta M S Lau. 2014. “Advances in Paediatric Cancer Treatment.” *Translational Pediatrics* 3 (2). AME Publications: 156–82. doi:10.3978/j.issn.2224-4336.2014.02.01.

Salwinski, L., Christopher S Miller, Adam J Smith, Frank K Pettit, James U Bowie, and David Eisenberg. 2004. “The Database of Interacting Proteins: 2004 Update.” *Nucleic Acids Research* 32 (90001): 449D–451. doi:10.1093/nar/gkh086.

Sanborn, J Zachary, Sofie R Salama, Mia Grifford, Cameron W Brennan, Tom Mikkelsen, Suresh Jhanwar, Sol Katzman, Lynda Chin, and David Haussler. 2013. “Double Minute Chromosomes in Glioblastoma Multiforme Are Revealed by Precise Reconstruction of Oncogenic Amplicons.” *Cancer Research* 73 (19). NIH Public Access: 6036–45. doi:10.1158/0008-5472.CAN-13-0186.

Saunders, C T, W S Wong, S Swamy, J Becq, L J Murray, and R K Cheetham. 2012. “Strelka: Accurate Somatic Small-Variant Calling from Sequenced Tumor-Normal Sample Pairs.” *Bioinformatics* 28 (14): 1811–17.

Sawyers, C L. 1999. “Chronic Myeloid Leukemia.” *N Engl J Med* 340 (17): 1330–40. doi:10.1056/NEJM199904293401706.

Schadendorf, Dirk, F. Stephen Hodi, Caroline Robert, Jeffrey S. Weber, Kim Margolin, Omid Hamid, Debra Patt, Tai-Tsang Chen, David M. Berman, and Jedd D. Wolchok. 2015. “Pooled Analysis of Long-Term Survival Data From Phase II and Phase III Trials of Ipilimumab in Unresectable or Metastatic Melanoma.” *Journal of Clinical Oncology* 33 (17): 1889–94. doi:10.1200/JCO.2014.56.2736.

Schmaußer, Bernd, Mindaugas Andrulis, Simon Endrich, Hans-Konrad Müller-Hermelink, and Matthias Eck. 2005. “Toll-like Receptors TLR4, TLR5 and

TLR9 on Gastric Carcinoma Cells: An Implication for Interaction with Helicobacter Pylori.” *International Journal of Medical Microbiology* 295 (3): 179–85. doi:10.1016/j.ijmm.2005.02.009.

Schneider, Günter, Marc Schmidt-Supprian, Roland Rad, and Dieter Saur. 2017. “Tissue-Specific Tumorigenesis: Context Matters.” *Nature Reviews Cancer* 17 (4). Nature Publishing Group: 239–53. doi:10.1038/nrc.2017.5.

Schölkopf, Bernhard., and Alexander J. Smola. 2002. *Learning with Kernels : Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press.

Schölkopf, Bernhard, John C. Platt, John Shawe-Taylor, Alex J. Smola, and Robert C. Williamson. 2001. “Estimating the Support of a High-Dimensional Distribution.” *Neural Computation* 13 (7). MIT Press: 1443–71. doi:10.1162/089976601750264965.

Schwarz, J M, C Rodelsperger, M Schuelke, and D Seelow. 2010. “MutationTaster Evaluates Disease-Causing Potential of Sequence Alterations.” *Nat Methods* 7 (8): 575–76. doi:10.1038/nmeth0810-575.

Secrier, Maria, Xiaodun Li, Nadeera De Silva, Matthew D. Eldridge, Gianmarco Contino, Jan Bornschein, Shona Macrae, et al. 2016. “Mutational Signatures in Esophageal Adenocarcinoma Define Etiologically Distinct Subgroups with Therapeutic Relevance.” *Nature Genetics* 48 (10): 1131–41. doi:10.1038/ng.3659.

Segditsas, S, A J Rowan, K Howarth, A Jones, S Leedham, N A Wright, P Gorman, et al. 2009. “APC and the Three-Hit Hypothesis.” *Oncogene* 28 (1): 146–55. doi:10.1038/onc.2008.361.

Sharma, S, T K Kelly, and P A Jones. 2010. “Epigenetics in Cancer.” *Carcinogenesis* 31 (1): 27–36. doi:10.1093/carcin/bgp220.

Shihab, H A, J Gough, D N Cooper, I N Day, and T R Gaunt. 2013. “Predicting the Functional Consequences of Cancer-Associated Amino Acid Substitutions.” *Bioinformatics* 29 (12): 1504–10.

Shimokawa, Takashi, Satoshi Matsushima, Takuya Tsunoda, Hideaki Tahara, Yusuke Nakamura, and Yoichi Furukawa. 2006. “Identification of TOMM34, Which Shows Elevated Expression in the Majority of Human Colon Cancers, as a Novel Drug Target.” *International Journal of Oncology* 29 (2). Spandidos Publications: 381–86. doi:10.3892/ijo.29.2.381.

Simpson, Andrew JG. 2009. “Sequence-Based Advances in the Definition of

Cancer-Associated Gene Mutations." *Current Opinion in Oncology* 21 (1): 47–52. doi:10.1097/CCO.0b013e32831de4b9.

Singh, Siddharth, Anamay N. Sharma, Mohammad Hassan Murad, Navtej S. Buttar, Hashem B. El–Serag, David A. Katzka, and Prasad G. Iyer. 2013. "Central Adiposity Is Associated With Increased Risk of Esophageal Inflammation, Metaplasia, and Adenocarcinoma: A Systematic Review and Meta-Analysis." *Clinical Gastroenterology and Hepatology* 11 (11): 1399–1412.e7. doi:10.1016/j.cgh.2013.05.009.

Sjoblom, T, S Jones, L D Wood, D W Parsons, J Lin, T D Barber, D Mandelker, et al. 2006. "The Consensus Coding Sequences of Human Breast and Colorectal Cancers." *Science* 314 (5797): 268–74. doi:10.1126/science.1133427.

Smith, G, R Bounds, H Wolf, R J Steele, F A Carey, and C R Wolf. 2010. "Activating K-Ras Mutations Outwith 'hotspot' Codons in Sporadic Colorectal Tumours - Implications for Personalised Cancer Medicine." *Br J Cancer* 102 (4): 693–703. doi:10.1038/sj.bjc.6605534.

Song, Qingxuan, Sofia D Merajver, and Jun Z Li. 2015. "Cancer Classification in the Genomic Era: Five Contemporary Problems." *Human Genomics* 9 (October). BioMed Central: 27. doi:10.1186/s40246-015-0049-8.

Soussi, T, and K G Wiman. 2015. "TP53: An Oncogene in Disguise." *Cell Death & Differentiation* 22 (8): 1239–49. doi:10.1038/cdd.2015.53.

Sparkes, R S, A L Murphree, R W Lingua, M C Sparkes, L L Field, S J Funderburk, and W F Benedict. 1983. "Gene for Hereditary Retinoblastoma Assigned to Human Chromosome 13 by Linkage to Esterase D." *Science* 219 (4587): 971–73. https://www.ncbi.nlm.nih.gov/pubmed/6823558.

Sparkes, R S, M C Sparkes, M G Wilson, J W Towner, W Benedict, A L Murphree, and J J Yunis. 1980. "Regional Assignment of Genes for Human Esterase D and Retinoblastoma to Chromosome Band 13q14." *Science* 208 (4447): 1042–44. https://www.ncbi.nlm.nih.gov/pubmed/7375916.

Steffen, Annika, José-Maria Huerta, Elisabete Weiderpass, H.Bas Bueno-de-Mesquita, Anne M. May, Peter D. Siersema, Rudolf Kaaks, et al. 2015. "General and Abdominal Obesity and Risk of Esophageal and Gastric Adenocarcinoma in the European Prospective Investigation into Cancer and Nutrition." *International Journal of Cancer* 137 (3): 646–57. doi:10.1002/ijc.29432.

Stephens, Philip J., Chris D. Greenman, Beiyuan Fu, Fengtang Yang, Graham R. Bignell, Laura J. Mudie, Erin D. Pleasance, et al. 2011. “Massive Genomic Rearrangement Acquired in a Single Catastrophic Event during Cancer Development.” *Cell* 144 (1): 27–40. doi:10.1016/j.cell.2010.11.055.

Stephens, Philip J., David J. McBride, Meng-Lay Lin, Ignacio Varela, Erin D. Pleasance, Jared T. Simpson, Lucy A. Stebbings, et al. 2009. “Complex Landscapes of Somatic Rearrangement in Human Breast Cancer Genomes.” *Nature* 462 (7276): 1005–10. doi:10.1038/nature08645.

Sterner, J M, S Dew-Knight, C Musahl, S Kornbluth, and J M Horowitz. 1998. “Negative Regulation of DNA Replication by the Retinoblastoma Protein Is Mediated by Its Association with MCM7.” *Molecular and Cellular Biology* 18 (5). American Society for Microbiology: 2748–57. doi:10.1128/MCB.18.5.2748.

Stratton, M R, P J Campbell, and P A Futreal. 2009. “The Cancer Genome.” *Nature* 458 (7239): 719–24. doi:10.1038/nature07943.

Stratton, Michael R. 2011. “Exploring the Genomes of Cancer Cells: Progress and Promise.” *Science (New York, N.Y.)* 331 (6024): 1553–58. doi:10.1126/science.1204040.

Suttorp, Meinolf, Philipp Schulze, Ingmar Glauche, Gudrun Göhring, Nils von Neuhoff, Markus Metzler, Petr Sedlacek, et al. 2018. “Front-Line Imatinib Treatment in Children and Adolescents with Chronic Myeloid Leukemia: Results from a Phase III Trial.” *Leukemia* 32 (7): 1657–69. doi:10.1038/s41375-018-0179-9.

Suykens, J.A.K. 2001. “Nonlinear Modelling and Support Vector Machines.” In *IMTC 2001. Proceedings of the 18th IEEE Instrumentation and Measurement Technology Conference. Rediscovering Measurement in the Age of Informatics (Cat. No.01CH 37188)*, 1:287–94. IEEE. doi:10.1109/IMTC.2001.928828.

Tamborero, D, A Gonzalez-Perez, and N Lopez-Bigas. 2013. “OncodriveCLUST: Exploiting the Positional Clustering of Somatic Mutations to Identify Cancer Genes.” *Bioinformatics* 29 (18): 2238–44. doi:10.1093/bioinformatics/btt395.

Tan, V. Y. F., and C. Fevotte. 2013. “Automatic Relevance Determination in Nonnegative Matrix Factorization with the /Spl Beta/-Divergence.” *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35 (7): 1592–

1605. doi:10.1109/TPAMI.2012.240.

Tate, John G, Sally Bamford, Harry C Jubb, Zbyslaw Sondka, David M Beare, Nidhi Bindal, Harry Boutselakis, et al. 2018. “COSMIC: The Catalogue Of Somatic Mutations In Cancer.” *Nucleic Acids Research*, October. doi:10.1093/nar/gky1015.

Teras, L R, D E Rollison, M Pawlita, A Michel, J Brozy, S de Sanjose, J L Blase, and S M Gapstur. 2015. “Epstein-Barr Virus and Risk of Non-Hodgkin Lymphoma in the Cancer Prevention Study-II and a Meta-Analysis of Serologic Studies.” *Int J Cancer* 136 (1): 108–16. doi:10.1002/ijc.28971.

The Cancer Genome Atlas. 2017. “Integrated Genomic Characterization of Oesophageal Carcinoma.” *Nature* 541 (7636). Macmillan Publishers Limited, part of Springer Nature. All rights reserved.: 169–75. doi:10.1038/nature20805.

Thrift, A. P., and D. C. Whiteman. 2012. “The Incidence of Esophageal Adenocarcinoma Continues to Rise: Analysis of Period and Birth Cohort Effects on Recent Trends.” *Annals of Oncology* 23 (12): 3155–62. doi:10.1093/annonc/mds181.

Thrift, Aaron P. 2016. “The Epidemic of Oesophageal Carcinoma: Where Are We Now?” *Cancer Epidemiology* 41 (April): 88–95. doi:10.1016/j.canep.2016.01.013.

Thrift, Aaron P., Nicholas J. Shaheen, Marilie D. Gammon, Leslie Bernstein, Brian J. Reid, Lynn Onstad, Harvey A. Risch, et al. 2014. “Obesity and Risk of Esophageal Adenocarcinoma and Barrett’s Esophagus: A Mendelian Randomization Study.” *JNCI: Journal of the National Cancer Institute* 106 (11). doi:10.1093/jnci/dju252.

Tian, Rui, Malay K Basu, and Emidio Capriotti. 2015. “Computational Methods and Resources for the Interpretation of Genomic Variants in Cancer.” *BMC Genomics* 16 Suppl 8 (Suppl 8). BioMed Central: S7. doi:10.1186/1471-2164-16-S8-S7.

Tibshirani, R., T. Hastie, B. Narasimhan, and G. Chu. 2002. “Diagnosis of Multiple Cancer Types by Shrunken Centroids of Gene Expression.” *Proceedings of the National Academy of Sciences* 99 (10): 6567–72. doi:10.1073/pnas.082099299.

Tokheim, C J, N Papadopoulos, K W Kinzler, B Vogelstein, and R Karchin. 2016. “Evaluating the Evaluation of Cancer Driver Genes.” *Proc Natl Acad*

Sci U S A 113 (50): 14330–35. doi:10.1073/pnas.1616440113.

Tomasetti, Cristian, Bert Vogelstein, and Giovanni Parmigiani. 2012. “Half or More of the Somatic Mutations in Cancers of Self-Renewing Tissues Originate Prior to Tumor Initiation.” *Proceedings of the National Academy of Sciences* 110: 1999–2004. doi:10.1073/pnas.1221068110.

Tomlinson, I, and W Bodmer. 1999. “Selection, the Mutation Rate and Cancer: Ensuring That the Tail Does Not Wag the Dog.” *Nat Med* 5 (1): 11–12. doi:10.1038/4687.

Tomlinson, I P, M R Novelli, and W F Bodmer. 1996. “The Mutation Rate and Cancer.” *Proc Natl Acad Sci U S A* 93 (25): 14800–803.

Tomlinson, I, P Sasieni, and W Bodmer. 2002. “How Many Mutations in a Cancer?” *Am J Pathol* 160 (3): 755–58. doi:10.1016/S0002-9440(10)64896-1.

Torkamani, Ali, Gennady Verkhivker, and Nicholas J. Schork. 2009. “Cancer Driver Mutations in Protein Kinase Genes.” *Cancer Letters* 281 (2). Elsevier: 117–27. doi:10.1016/J.CANLET.2008.11.008.

Trevellin, Elisabetta, Marco Scarpa, Amedeo Carraro, Francesca Lunardi, Andromachi Kotsafti, Andrea Porzionato, Luca Saadeh, et al. 2015. “Esophageal Adenocarcinoma and Obesity: Peritumoral Adipose Tissue Plays a Role in Lymph Node Invasion.” *Oncotarget* 6 (13). Impact Journals, LLC: 11203–15. doi:10.18632/oncotarget.3587.

Tsoumakas, Grigorios, Grigorios Tsoumakas, and Ioannis Katakis. 2007. “Multi-Label Classification: An Overview.” *INT J DATA WAREHOUSING AND MINING* 2007: 1--13. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.104.9401.

Tsuruoka, Yoshimasa, ‡ Jun’ichi Tsujii, and Sophia Ananiadou. 2009. “Stochastic Gradient Descent Training for L1-Regularized Log-Linear Models with Cumulative Penalty.” In *Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP,* 477–85. https://www.aclweb.org/anthology/P/P09/P09-1054.pdf.

Vandin, F, E Upfal, and B J Raphael. 2011. “Algorithms for Detecting Significantly Mutated Pathways in Cancer.” *J Comput Biol* 18 (3): 507–22. doi:10.1089/cmb.2010.0265.

Verweij, Jaap, Paolo G Casali, John Zalcberg, Axel LeCesne, Peter Reichardt, Jean-Yves Blay, Rolf Issels, et al. 2004. “Progression-Free Survival in

Gastrointestinal Stromal Tumours with High-Dose Imatinib: Randomised Trial." *The Lancet* 364 (9440): 1127–34. doi:10.1016/S0140-6736(04)17098-0.

Villanueva, Augusto, Anna Portela, Sergi Sayols, Carlo Battiston, Yujin Hoshida, Jesús Méndez-González, Sandrine Imbeaud, et al. 2015. "DNA Methylation-Based Prognosis and Epidrivers in Hepatocellular Carcinoma." *Hepatology* 61 (6): 1945–56. doi:10.1002/hep.27732.

Vogel, Charles L., Melody A. Cobleigh, Debu Tripathy, John C. Gutheil, Lyndsay N. Harris, Louis Fehrenbacher, Dennis J. Slamon, et al. 2002. "Efficacy and Safety of Trastuzumab as a Single Agent in First-Line Treatment of *HER2* -Overexpressing Metastatic Breast Cancer." *Journal of Clinical Oncology* 20 (3): 719–26. doi:10.1200/JCO.2002.20.3.719.

Vogelstein, B, and K W Kinzler. 2004. "Cancer Genes and the Pathways They Control." *Nat Med* 10 (8): 789–99. doi:10.1038/nm1087.

Vogelstein, B, N Papadopoulos, V E Velculescu, S Zhou, L A Diaz Jr., and K W Kinzler. 2013. "Cancer Genome Landscapes." *Science* 339 (6127): 1546–58. doi:10.1126/science.1235122.

Vogt, P K. 2012. "Retroviral Oncogenes: A Historical Primer." *Nat Rev Cancer* 12 (9): 639–48. doi:10.1038/nrc3320.

Voutilainen, Markku, Pentti Sipponen, Jukka-Pekka Mecklin, Matti Juhola, and Martti Färkkilä. 2000. "Gastroesophageal Reflux Disease: Prevalence, Clinical, Endoscopic and Histopathological Findings in 1,128 Consecutive Patients Referred for Endoscopy Due to Dyspeptic and Reflux Symptoms." *Digestion* 61 (1): 6–13. doi:10.1159/000007730.

Wagner, Meike, Michael Koslowski, Claudia Paret, Marcus Schmidt, Ozlem Türeci, and Ugur Sahin. 2013. "NCOA3 Is a Selective Co-Activator of Estrogen Receptor α-Mediated Transactivation of PLAC1 in MCF-7 Breast Cancer Cells." *BMC Cancer* 13 (December). BioMed Central: 570. doi:10.1186/1471-2407-13-570.

Walther, A, R Houlston, and I Tomlinson. 2008. "Association between Chromosomal Instability and Prognosis in Colorectal Cancer: A Meta-Analysis." *Gut* 57 (7): 941–50. doi:10.1136/gut.2007.135004.

Wang, Hongbing, Yanqi Shi, Xuan Zhou, Qianzhao Zhou, Shizhi Shao, and Athman Bouguettaya. 2010. "Web Service Classification Using Support Vector Machine." doi:10.1109/ICTAI.2010.9.

Wang, K, M Li, and H Hakonarson. 2010. “{ANNOVAR}: Functional Annotation of Genetic Variants from High-Throughput Sequencing Data.” *Nucleic Acids Res.* 38 (16): e164.

Wang, Z., T. G. Da Silva, K. Jin, X. Han, P. Ranganathan, X. Zhu, A. Sanchez-Mejias, et al. 2014. “Notch Signaling Drives Stemness and Tumorigenicity of Esophageal Adenocarcinoma.” *Cancer Research* 74 (21): 6364–74. doi:10.1158/0008-5472.CAN-14-2051.

Watson, J D, and F H Crick. 1953. “Molecular Structure of Nucleic Acids; a Structure for Deoxyribose Nucleic Acid.” *Nature* 171 (4356): 737–38. http://www.ncbi.nlm.nih.gov/pubmed/13054692.

Weaver, Jamie M. J., Caryn S. Ross-Innes, and Rebecca C. Fitzgerald. 2014. “The ‘–Omics’ Revolution and Oesophageal Adenocarcinoma.” *Nature Reviews Gastroenterology & Hepatology* 11 (1). Nature Publishing Group: 19–27. doi:10.1038/nrgastro.2013.150.

Weaver, Jamie M J, Caryn S Ross-Innes, Nicholas Shannon, Andy G Lynch, Tim Forshew, Mariagnese Barbera, Muhammed Murtaza, et al. 2014. “Ordering of Mutations in Preinvasive Disease Stages of Esophageal Carcinogenesis.” *Nat. Genet.* 46 (8). Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.: 837–43.

Weir, Barbara, Xiaojun Zhao, and Matthew Meyerson. 2004. “Somatic Alterations in the Human Cancer Genome.” *Cancer Cell* 6 (5). Cell Press: 433–38. doi:10.1016/J.CCR.2004.11.004.

Weiske, J., K. F. Albring, and O. Huber. 2007. “The Tumor Suppressor Fhit Acts as a Repressor of -Catenin Transcriptional Activity.” *Proceedings of the National Academy of Sciences* 104 (51): 20344–49. doi:10.1073/pnas.0703664105.

Wolf, R. M., Nicole Draghi, Xiquan Liang, Chengkai Dai, Lene Uhrbom, Charlotta Eklöf, Bengt Westermark, Eric C Holland, and Marilyn D Resh. 2003. “P190RhoGAP Can Act to Inhibit PDGF-Induced Gliomas in Mice: A Putative Tumor Suppressor Encoded on Human Chromosome 19q13.3.” *Genes & Development* 17 (4): 476–87. doi:10.1101/gad.1040003.

Woo, Janghee, Stacey A Cohen, and Jonathan E Grim. 2015. “Targeted Therapy in Gastroesophageal Cancers: Past, Present and Future.” *Gastroenterol. Rep.* 3 (4): 316–29.

Wood, L D, D W Parsons, S Jones, J Lin, T Sjoblom, R J Leary, D Shen, et al.

2007. “The Genomic Landscapes of Human Breast and Colorectal Cancers.” *Science* 318 (5853): 1108–13. doi:10.1126/science.1145720.

Xiao, F., Z. Zuo, G. Cai, S. Kang, X. Gao, and T. Li. 2009. “MiRecords: An Integrated Resource for MicroRNA-Target Interactions.” *Nucleic Acids Research* 37 (Database): D105–10. doi:10.1093/nar/gkn851.

Xue, Yuan, and William R Wilcox. 2016. “Changing Paradigm of Cancer Therapy: Precision Medicine by next-Generation Sequencing.” *Cancer Biology & Medicine* 13 (1). Chinese Anti-Cancer Association: 12–18. doi:10.28092/j.issn.2095-3941.2016.0003.

Yaghmour, George, Manjari Pandey, Catherine Ireland, Kruti Patel, Sara Nunnery, Daniel Powell, Scott Baum, Eric Wiedower, Lee S Schwartzberg, and Michael G Martin. 2016. “Role of Genomic Instability in Immunotherapy with Checkpoint Inhibitors.” *Anticancer Research* 36 (8). International Institute of Anticancer Research: 4033–38. http://www.ncbi.nlm.nih.gov/pubmed/27466509.

Yang, Xiang-Jiao, and Edward Seto. 2008. “Lysine Acetylation: Codified Crosstalk with Other Posttranslational Modifications.” *Molecular Cell* 31 (4): 449–61. doi:10.1016/j.molcel.2008.07.002.

Ye, ZhenLong, HuaJun Jin, and QiJun Qian. 2015. “Argonaute 2: A Novel Rising Star in Cancer Research.” *Journal of Cancer* 6 (9). Ivyspring International Publisher: 877–82. doi:10.7150/jca.11735.

Yoshida, K, and I Inoue. 2004. “Regulation of Geminin and Cdt1 Expression by {E2F} Transcription Factors.” *Oncogene* 23 (21): 3802–12.

Young, Kate, and Ian Chau. 2016. “Targeted Therapies for Advanced Oesophagogastric Cancer: Recent Progress and Future Directions.” *Drugs* 76 (1): 13–26.

Zhang, Cheng-Zhong, Mitchell L Leibowitz, and David Pellman. 2013. “Chromothripsis and beyond: Rapid Genome Evolution from Complex Chromosomal Rearrangements.” *Genes & Development* 27 (23). Cold Spring Harbor Laboratory Press: 2513–30. doi:10.1101/gad.229559.113.

Zhang, Shaoyi, M. Maruf Hossain, Md. Rafiul Hassan, James Bailey, and Kotagiri Ramamohanarao. 2009. “Feature Weighted SVMs Using Receiver Operating Characteristics.” In *Proceedings of the 2009 SIAM International Conference on Data Mining*, 497–508. Philadelphia, PA: Society for Industrial and Applied Mathematics. doi:10.1137/1.9781611972795.43.

Zhang, W, R Hong, L Xue, Y Ou, X Liu, Z Zhao, W Xiao, et al. 2017. “Piccolo Mediates EGFR Signaling and Acts as a Prognostic Biomarker in Esophageal Squamous Cell Carcinoma.” *Oncogene* 36 (27): 3890–3902.